universität wien

BAK3: Introduction to Quantitative Methods

Week 13: Inference in Regression

Leonardo Carella

universität wien

The Plan for today

- Statistics:
  - Recap: Inference + Multiple Regression.
  - Standard error, confidence interval, t-statistic and *p*-value of regression coefficients.
- Coding in R:
  - Visualising Regression Models with Uncertainty

universität wien

Inference Recap

- We observe a **sample estimate** (a sample mean or a difference-in-means). How does it relate to the **population parameter**?
- Measures of uncertainty:
  - **Standard Error**: estimated std. deviation of the estimate across repeated sampling from the population.
  - **95% Confidence Interval**: range of values calculated in such a way that in 95% of the samples it will include the population parameter.

universität wien

Inference Recap

- We observe a **sample estimate** (a sample mean or a difference-in-means). How strong is the evidence against the null hypothesis that the population parameter equals a given value (e.g. zero)?
- Hypothesis testing:
  - ***t*-statistic**: difference between the estimate and the parameter under the null hypothesis, expressed in number of standard error.
  - ***p*-value**: the probability of observing a sample estimate at least as extreme as the one we observe, under the null hypothesis.

universität wien

Inference Recap

- These things are related to each other. E.g., for a sample mean $\bar{x}$…
- $\text{SE}_{\bar{x}} = \dfrac{s}{\sqrt{n}}$
- $\text{C.I.}_{\mu}(95\%) = \bar{x} \pm 1.96 \times \text{SE}_{\bar{x}}$
- t-statistic $= \dfrac{\bar{x} - \mu_0}{\text{SE}_{\bar{x}}}$
- Each t-statistic has an associated p-value (there's a formula for it, but we didn't learn it — which is fine, as long as you get the intuition).

universität wien

Regression Recap

- Our model of reality:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_p X_p + \varepsilon$$

- Where each $\beta_1, \beta_2, \beta_3 \ldots$ represents the predicted change in Y associated with a one-unit increase in $X_1, X_2, X_3 \ldots$ respectively **holding covariates constant**.

- How do we pick the α and βs? **Ordinary Least Squares**: the set of coefficients that minimise the sum of squared residuals.

universität wien

Regression Recap

- Interpretation:
- Numerical variables: “a one-unit increase in X is associated with a β increase in Y, **holding covariates constant**.”
- Categorical variables: “the predicted difference between observations in category k **and the reference category** is β, holding covariates constant.”
- Usually, we’re interested in interpreting substantively only the **slope** for **one variable**.

	Dependent variable:
	Life Satisfaction (0–10)
Age	0.013*** (0.004)
Income Decile	0.163*** (0.019)
Female	0.288*** (0.100)
Religiosity (0–10)	0.022 (0.017)
Years of Education	−0.003 (0.014)
Divorced	−0.354 (0.299)
Single	−0.118 (0.131)
Widowed	−0.412** (0.189)
Constant	5.713*** (0.321)
Observations	1,601
R^2	0.078
Adjusted R^2	0.073
Note:	*p<0.1; **p<0.05; ***p<0.01

universität wien

Inference in Regressions

- Regression as a model of the data-generating process (DGP):

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_p X_p + \varepsilon$$

- We are assuming that Y is linearly related to our Xs plus some random error ε: some observations will have a higher value of Y than predicted, some lower, at random. What does it mean that there's some random error?

- That, if we could repeat the same DGP, we would get different values of Y.

- Therefore, in a regression, we can treat our observed data as one specific realisation of the DGP: that is, as a **sample from a hypothetical population.**

universität wien

Inference in Regressions

- In this sense, (OLS) linear regression is an estimator, like the sample mean or difference-in-means etc.
- The only difference is that in this case we get more than one estimate ($\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$…) of more than one parameter (α, β_1, β_2…) of the ‘population’ regression (which is an abstract model).
- That’s why our estimates (and we’re generally interested in our $\hat{\beta}$s) come with similar measures of **uncertainty** as our sample mean, difference-in-means etc. from weeks 7 and 8.

Standard Errors of coefficients

	Intercept	Slope
Population	**−4.03**	**0.80**
Sample 1	−3.46	0.82
Sample 2	−1.71	0.58
Sample 3	−2.44	0.63
Sample 4	−4.98	0.90
Mean of Sample Estimates	⇝ −4.03	⇝ 0.80
Std. Dev of Sample Estimates	SE(α)	SE(β)

universität wien

Standard Errors of coefficients

- The Standard Error of a regression coefficient is the standard deviation of the coefficient across hypothetical repeated random sampling from the population.
- The distribution of these estimates is going to a normal distribution (thanks, Central Limit Theorem!)
- It expresses the uncertainty of the estimated coefficient. We're normally interested in the uncertainty of our $\hat{\beta}$s.
- Like the standard error of a sample mean, we can approximate the SE of a regression coefficient from a **single sample**.

universität wien

Standard Errors of coefficients

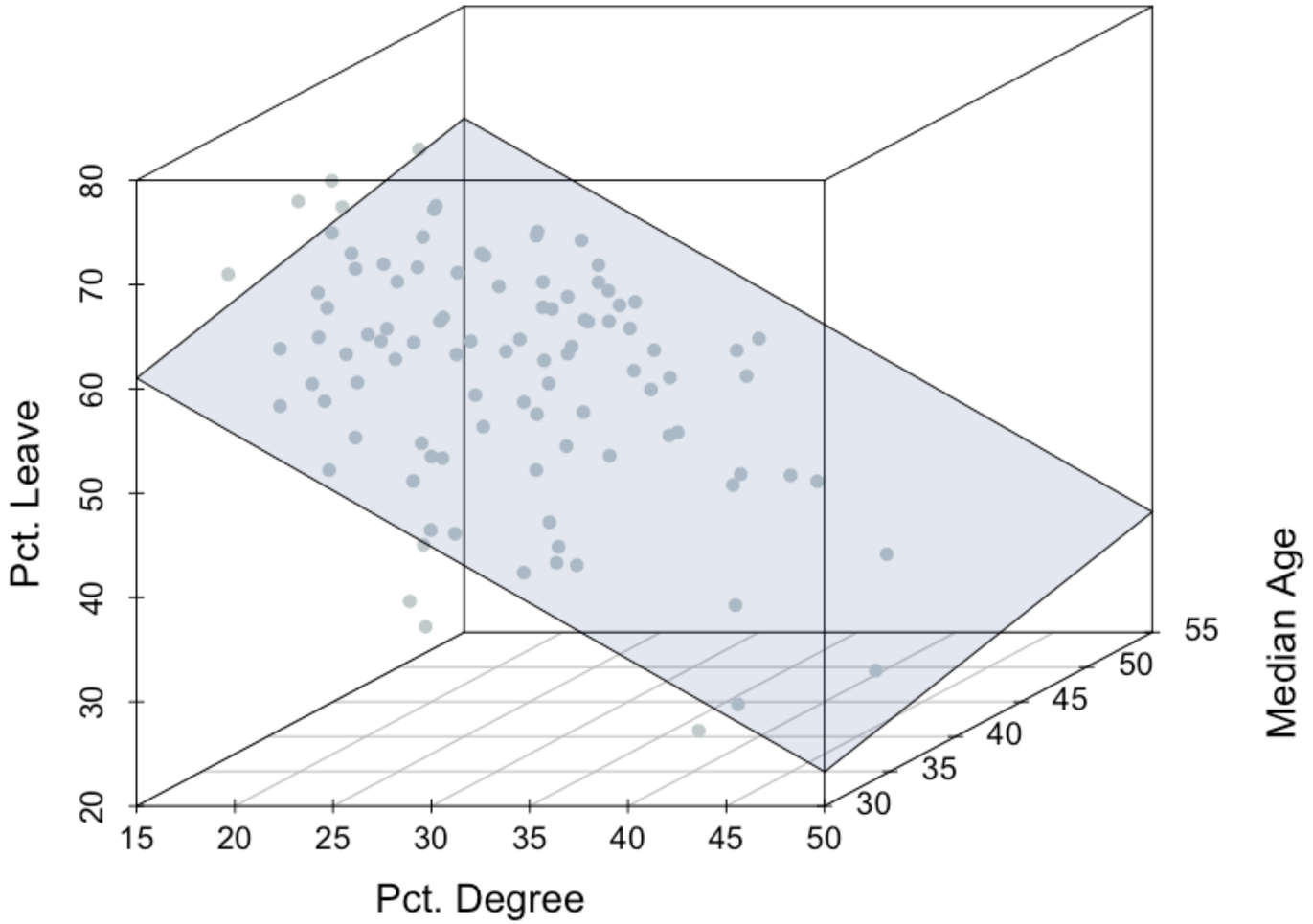
- Under the assumption that errors 'behave nicely', in a **bivariate** linear regression…

$$\text{S.E.}(\hat{\beta}) = \sqrt{\frac{\sum \hat{\varepsilon}^2}{(\frac{1}{n-2}) \sum (x_i - \bar{x})^2}}$$

- Your standard errors will be larger if…
  - X does a poor job at predicting Y $\rightarrow \sum \hat{\varepsilon}^2$, the sum of squared residuals, goes up.
  - X does not vary much $\rightarrow \sum (x_i - \bar{x})^2$ goes down.
  - Your sample is small $\rightarrow \frac{1}{n-2}$ goes down.

Generalising to Multiple Regression

- Same story, more complex math:
- Each variable ($X_1, X_2 \ldots$) will have an associated coefficient ($\hat{\beta}_1, \hat{\beta}_2$), which in turn come with their own standard error.
- Std. Error of $\hat{\beta}_2 \rightarrow$ estimated std. deviation of the slope of X_2 across repeated sampling from a hypothetical population.



universität wien

Standard Errors in R

```
> model_brexit <- lm(percent_leave ~ percent_degree +
median_age, data = brexit)
> summary(model_brexit)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    65.56026    3.36364  19.491  < 2e-16 ***
percent_degree -1.00905    0.04369 -23.097  < 2e-16 ***
median_age      0.34969    0.06941   5.038 7.31e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

universität wien

Standard Errors in R

```
> model_brexit <- lm(percent_leave ~ percent_degree +
median_age, data = brexit)
> stargazer(model_brexit, type = "text", single.row = TRUE)

================================================
                         Dependent variable:
                     ---------------------------
                            percent_leave
------------------------------------------------
percent_degree            -1.009*** (0.044)
median_age                0.350*** (0.069)
Constant                  65.560*** (3.364)
------------------------------------------------
Observations                     380
R2                              0.623
Adjusted R2                     0.621
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

universität wien

Confidence Intervals

- Identical idea as the confidence interval of a sample mean: it's a plausible range of values of the population parameter. Same formula:

$$\text{C.I.}_{0.95}(\beta) = \hat{\beta} \pm 1.96 \times \text{SE}(\hat{\beta})$$

- Why is it a plausible range of values? Because in 95% of the samples, the 95% confidence interval of the estimated slope **calculated this way** will include the population slope.

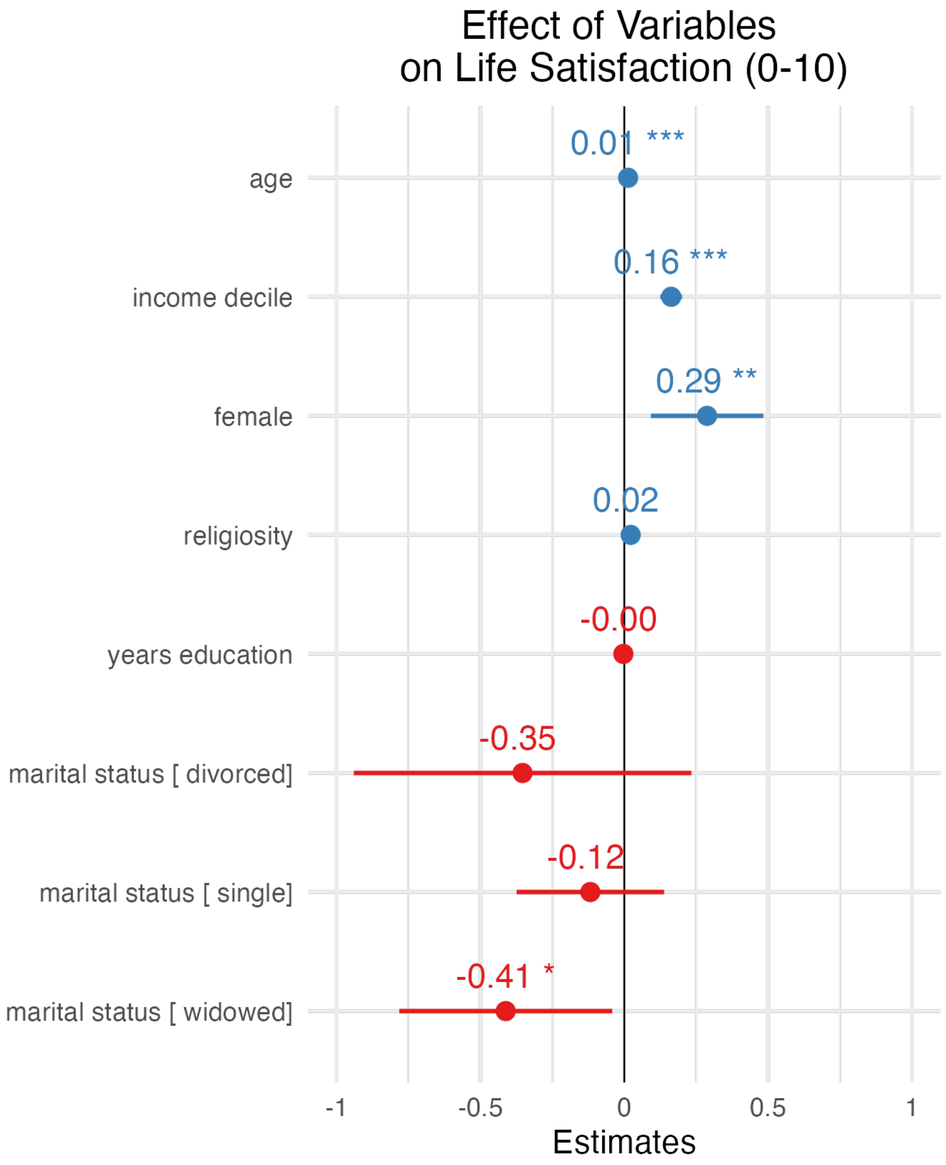
- Why does it work? Because the sampling distribution of β is normal.

Confidence Intervals

	Slope	S.E.	95% C.I	Includes β = 0.80?
Population	**0.80**			
Sample 1	0.82	0.12	(0.59 –1.06)	Yes
Sample 2	0.58	0.09	(0.39–0.77)	No
Sample 3	0.63	0.13	(0.37–0.89)	Yes
Sample 4	0.90	0.13	(0.64–1.15)	Yes
Over many repeated samples…	⇝ 0.80			In 95% (19 out of 20) samples

Visualising Regression Coefficients (1)

- **Coefficient Plot**: plots the estimated slopes with their confidence intervals.
- One potential drawback: predictors may be on very different scales, so not directly comparable.
- (Althout we're generally only interested in one X, so these comparisons are not that important)



universität wien

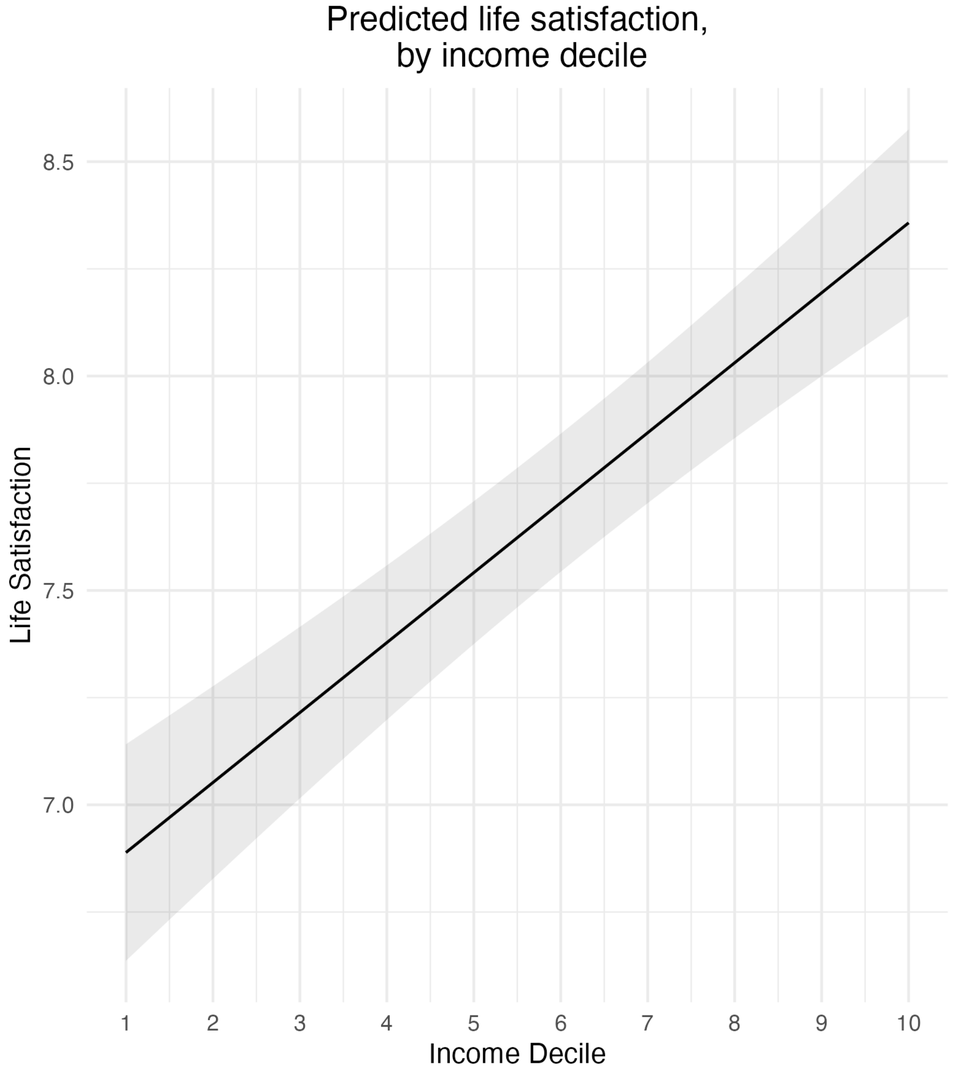
Visualising Regression Coefficients (2)

- **Standardised Coefficient Plot**: re-scales Xs and Y so that they have a standard deviation of 1.
- Plots $\hat{\beta}$ with confidence intervals: these are change in std. deviations in Y associated with one std. deviation increase in X.
- Drawback: categorical variables make little sense.

universität wien

Visualising Regression Coefficients (3)

- **Predicted values plot**: focus on a single independent variable X_1, hold all others constant at some "typical" value (e.g. the median or the mode).
- Plots $\hat{Y}$ with the confidence intervals **of the prediction**, i.e. plausible values of Y.
- Shows how $\hat{Y}$ varies as X_1 varies, while X_2, X_3, X_4 etc. are held constant at some "typical" value.



Predicted life satisfaction (0-10 scale) for a 51 year-old married woman with religiosity = 4/10 and 14 years of education

universität wien

Hypothesis Testing

- Commonly, we use regressions to estimate the relationship between X and Y, expressed by the slope.
- But if — as we assume — the data is generated with some random noise, how can we be sure that the relationship we found isn't due to the "random error" part of our model?
- In other words, how can we be sure that our sample slope estimate isn't "zero" (i.e. there's no relationship) in the hypothetical population?
- With a t-test! We test against the null hypothesis that $\beta = 0$.

universität wien

Hypothesis Testing

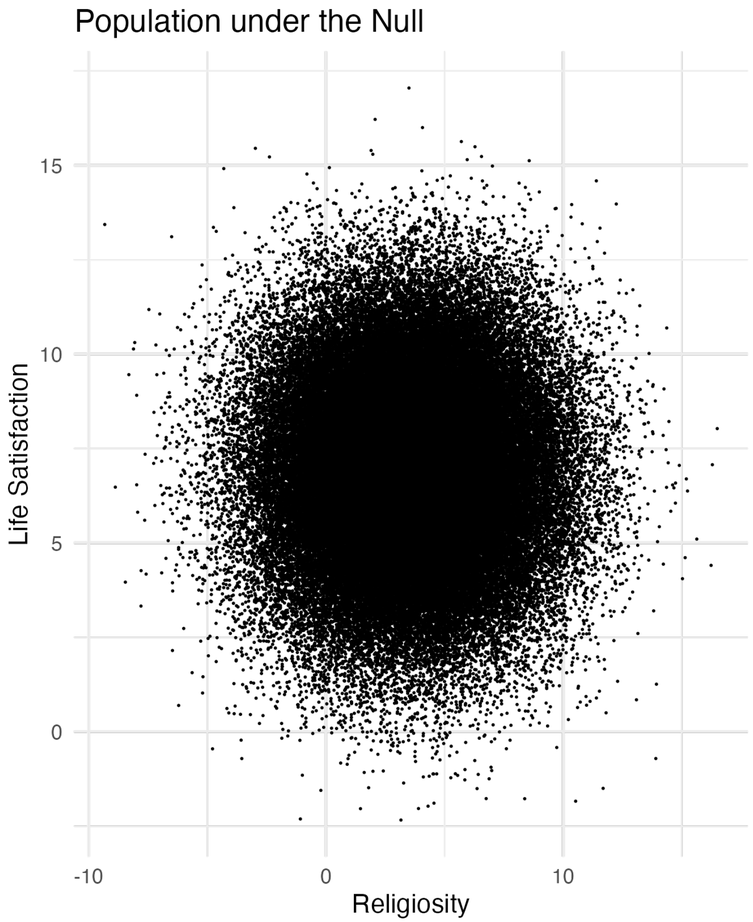
```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.5526     0.2173  30.150   <2e-16 ***
## religiosity    0.1053     0.0471   2.236   0.0262 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- From a sample of size n, we estimate $\hat{\beta} = 0.1053$.
- Assume a population where X and Y are completely uncorrelated, $Y_i = \alpha + 0X + \epsilon_i$ with normally distributed error (another assumption, but let's roll with it for now).
- If the 'true' $\beta = 0$, how likely is it that, over many samples of size n, we get **a slope as extreme** as $\hat{\beta}$? (i.e. $\hat{\beta}_s > 0.1053$ or $\hat{\beta}_s < -0.1053$)

universität wien

Hypothesis Testing

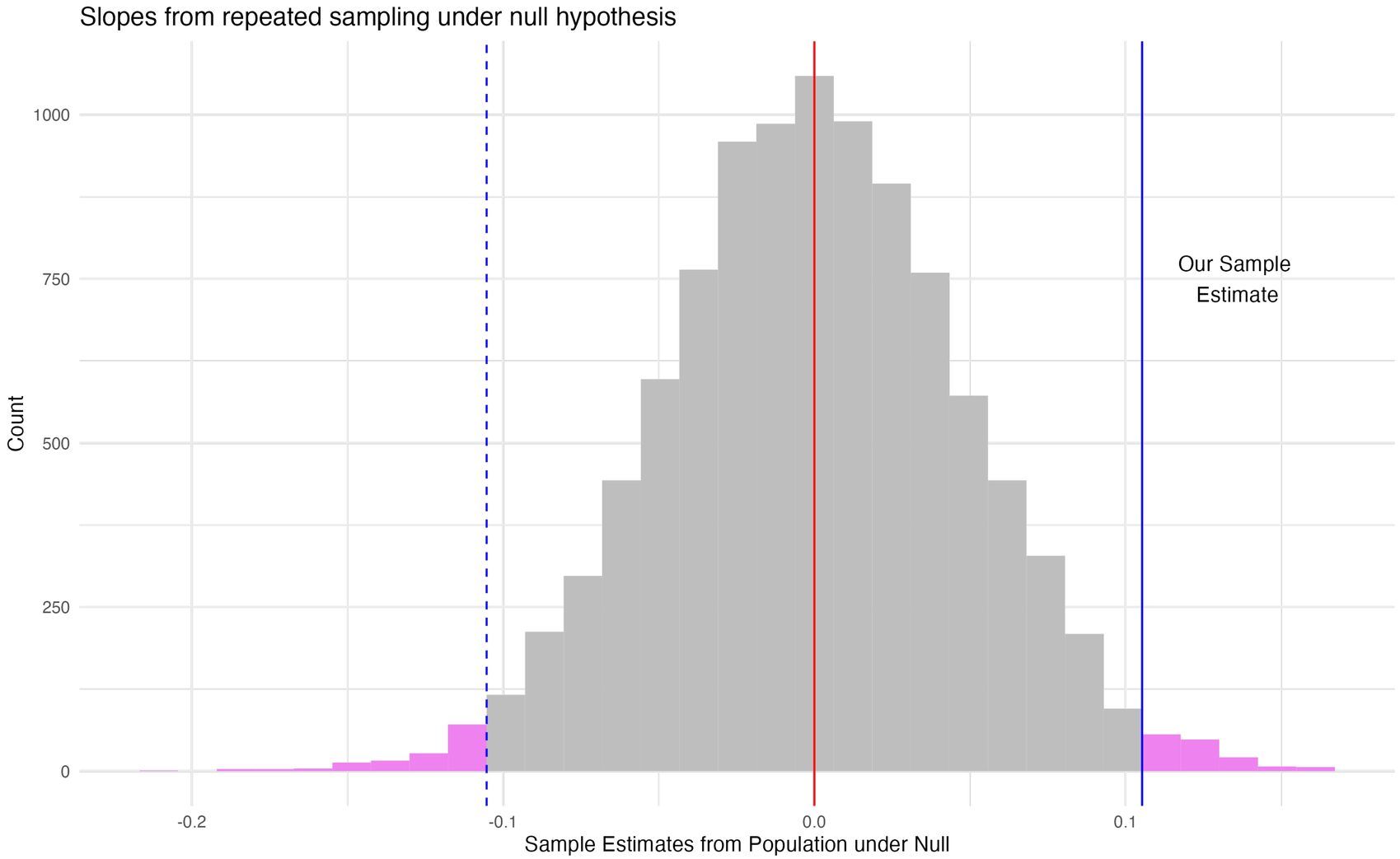
$$\text{Life Satisfaction} = \alpha + \beta\ \text{Religiosity} + \epsilon$$



	Slope
Our Data	**0.105**
Population under the null	**0**
Sample 1 from pop.	0.019
Sample 2 from pop.	0.042
Sample 3 from pop.	-0.011
Mean Over many repeated samples…	**⇝ 0**
Std. deviation of estimates over many repeated samples	**≈ SE(β)**

universität wien

Hypothesis Testing



universität wien

Hypothesis Testing

$$\text{t-statistic}(\hat{\beta}) = \frac{\hat{\beta} - \beta \text{ under the null}}{SE(\hat{\beta})} = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

- If $|t| \geq 1.96$, we reject the null (that there's no relationship between X and Y) at the 95% confidence level. In particular, this gives us confidence in the **direction** of the relationship (positive or negative).

- If $|t| < 1.96$, we **fail to reject the null**. Our relationship could plausibly have arisen from random error under the null hypothesis that $\beta = 0$, so we can't conclude that there's a positive or negative relationship between X and Y.

universität wien

Hypothesis Testing

- The **p-value** summarises our evidence against the null hypothesis, just like the t-statistic, and just like the p-value for a difference-in-mean etc.
- It's the probability of observing a **t-statistic** (and therefore an estimate) at least as extreme as one we observe, under the null hypothesis.
- A p-value below 0.05 means we reject the null at the 95% confidence level. Below 0.01, we reject the null at the 99% confidence level, and so on.
- It's ❌🔥💀❌🔥💀❌🔥💀❌ is true.

universität wien

In R...

We basically never care about the p-value of the intercept — we care about the uncertainty of our β!

```
summary(model1)
```

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      81.6906     2.8560   28.60   <2e-16 ***
## percent_degree  -1.0982     0.1063  -10.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Coefficients:
##                 Estimate Std. Error t value             Pr(>|t|)
## (Intercept)      81.6906     2.8560   28.60 <0.0000000000000002 ***
## percent_degree  -1.0982     0.1063  -10.33 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

universität wien

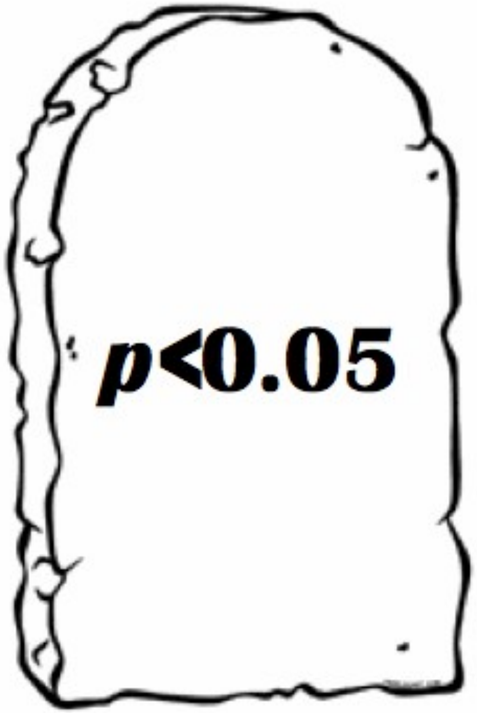
- Interpretation:
- “The relationship between age and life satisfaction is positive and significant at the 99% level ($p < 0.01$), holding covariates constant”
- “Holding covariates constant, people who are widowed have a significantly lower level of life satisfaction than people who are married ($p < 0.05$).”
- “Education is not significantly associated with life satisfaction in this model at conventional significance levels ($p > 0.05$).

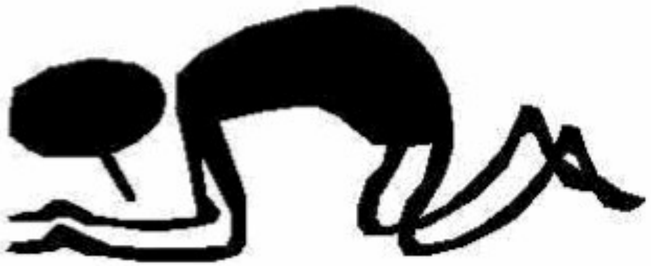
```
> stargazer(model9, type = "text", single.row = TRUE)

================================================
                         Dependent variable:
                     ---------------------------
                              life_satisf
------------------------------------------------
age                          0.013*** (0.004)
income_decile                0.163*** (0.019)
female                       0.288*** (0.100)
religiosity                   0.022 (0.017)
years_education              -0.003 (0.014)
marital_status divorced      -0.354 (0.299)
marital_status single        -0.118 (0.131)
marital_status widowed       -0.412** (0.189)
Constant                     5.713*** (0.321)
------------------------------------------------
Observations                        1,601
R2                                  0.078
Adjusted R2                         0.073
Residual Std. Error           1.947 (df = 1592)
F Statistic                16.778*** (df = 8; 1592)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Statistical Significance: Warnings

- Newcomers to statistics love **over-interpreting** measures of statistical significance like the *p*-value (*especially* the *p*-value):

- "The relationship is significant at the 99.99% level, so it's likely **true/causal/worth caring about**."





universität wien

Don't Be This Guy

"In 2020, Biden's tabulated votes (2,474,507) were much greater than Clinton's in 2016. […] I tested the hypothesis that the performance of the two Democrat [sic] candidates were statistically similar by comparing Clinton to Biden. […] I use the calculated Z-score to determine the p-value […]. This value corresponds to a confidence that I can reject the hypothesis many times more than one in a quadrillion times that the two outcomes were similar."

(Charles Cicchetti, Lawsuit filed by the State of Texas)

universität wien

Statistical Significance: Warnings

1. Your p-value is only as good as your model: no use if you have a bad model of reality (like the guy who thinks that there's nothing else that might explain the difference in vote percentage for two different candidates, in two different elections, 4 years apart).

2. You will get 'lucky' and find $p < 0.05$ one in twenty times if you regress nonsense on nonsense. Beware of fishing.

3. Cutoffs are arbitrary (and bad for science): $p = 0.049$ is just as good as $p = 0.051$.

4. Non-significant findings are valuable. Especially if we can be very confident about the fact that there's probably no meaningful relationship ('precise null').

5. Our assumptions over what it means for errors to be random are generally unrealistic. Advanced techniques for more "realistic" standard errors usually return larger p-values.

universität wien

Summing Up…

- In a regression model, our relationships are estimated with uncertainty, which we can think of as repeated sampling from a hypothetical population.
- Standard error and confidence interval express the uncertainty of our estimate. We are normally interested in the uncertainty of the slopes.
- P-value and t-statistics summarise our evidence against the null that the “real” parameter is zero in the (hypothetical) population from which our data is generated.
- In the case of the $\hat{\beta}$s, we are testing against the null that there’s no relationship (i.e. $\beta = 0$ in the population). If we can reject the null, we can be more confident in the direction (positive or negative) of the relationship we found.

universität wien

Next time

- Research design and linear regression: what things should I control for?
- Your seminar paper: expectations and tips.
- Revision of anything that may not be clear yet.
- And now, let's open RStudio…