Quantitative Economics: Regression

Regression with a Sample

Ian Crawford

Department of Economics

Nuffield College

Trinity Term, 2018

Weeks 3 & 4

Regression

1. Regression with the population
2. Regression with a sample
3. Regression and causal inference.

Regression with a sample

- Even in the era of Big Data we rarely have access to “the population”.
- We typically have to work with a random sample drawn from that population.
- Anything which we calculate from random samples is a statistic and subject to sampling variation.
- From these sample statistics we need to make statistical inferences about the population of interest.

Regression with a sample

Regression with a sample

- Key statistical ideas
- Confidence intervals
- t tests
- p values
- R-squared and adjusted R-squared
- F tests

Key statistical ideas

Analogue Estimators

To estimate a feature of the population (the “estimand”) use the corresponding feature of the sample (the “estimator”).

Law of Large Numbers

The sample mean converges to the population mean as the sample size grows.

Central Limit Theorem

The limiting distribution of the standardised mean is $N(0, 1)$.

The Population and the Sample LRM

Estimate the sample LRM by calculating sample analogues of the population moments.

The Population LRM

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X + u \\
\beta_0 &= \mathbb{E}\left[Y\right] - \beta_1 \mathbb{E}\left[X\right] \\
\beta_1 &= \frac{Cov(Y, X)}{Var(X)}
\end{aligned}
$$

The Sample LRM

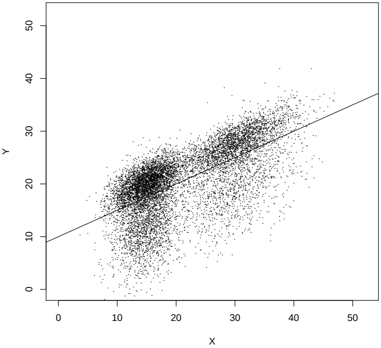
$$
\begin{aligned}
Y &= \hat{\beta}_0 + \hat{\beta}_1 X + \hat{u} \\
\hat{\beta}_0 &= \bar{Y} - \beta_1 \bar{X} \\
\hat{\beta}_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}
\end{aligned}
$$

Properties of the OLS estimator

If the sample size is sufficiently large then

- by the CLT, the sampling distributions of $\hat{\beta}_0, \hat{\beta}_1$ are approximately Normal;
- by the LLN, $\hat{\beta}_0, \hat{\beta}_1$ will be close to their true population values β_0, β_1 with high probability.

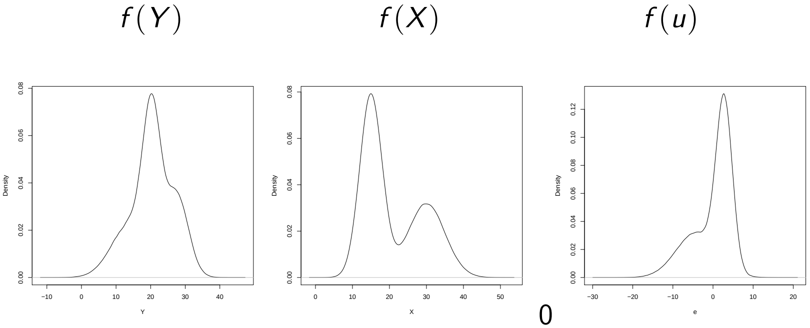
Properties of the OLS estimator- Normality



- This shows a scatterplot of the population of two variables $\{Y, X\}$.
- There are 1 million members of this population.
- The line is the population LRM

$$Y = 10 + \frac{1}{2}X + u$$

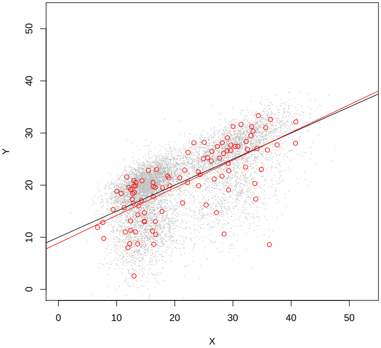
Properties of the OLS estimator- Normality



$f(Y)$ $f(X)$ $f(u)$

These are the densities of data $\{Y, X\}$ and the OLS residual u.
Plainly none of these are Normal.

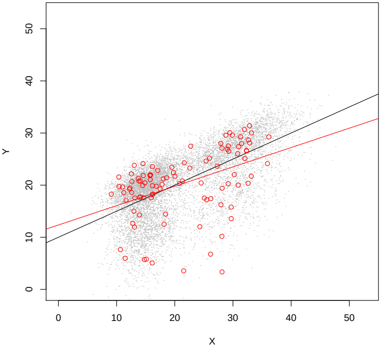
Properties of the OLS estimator - Normality



- This shows a sample of 100 observations from the population.
- The line is the sample/estimated LRM

$$Y = 8.89 + 0.53X + u$$

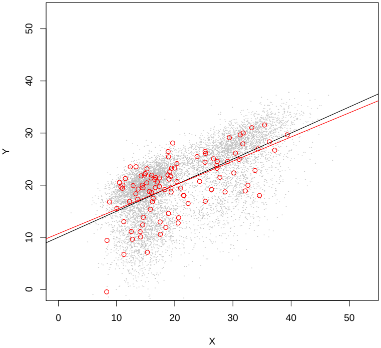
Properties of the OLS estimators - Normality



- This shows another sample of 100 observations from the population.
- The line is the sample/estimated LRM

$$Y = 12.36 + 0.37X + u$$

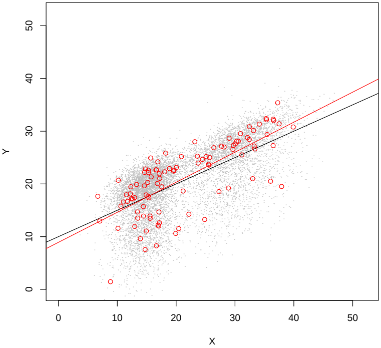
Properties of the LRM estimators - Normality



- This shows another sample of 100 observations from the population.
- The line is the sample/estimated LRM

$$Y = 10.7 + 0.46X + e$$

Properties of the LRM estimators - Normality



- This shows another sample of 100 observations from the population.
- The line is the sample/estimated LRM

$$Y = 8.92 + 0.57X + u$$

Properties of the OLS estimator - Normality

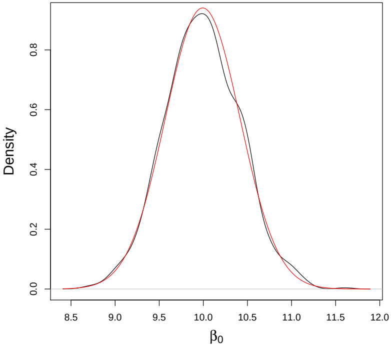
- We repeated the exercise of re-sampling and re-estimating four times.
- This simulates the sampling variation.
- We found that the coefficients varied a little each time:

$$\hat{\beta}_0 = \{8.89, 12.36, 10.7, 8.92, ...\}$$

$$\hat{\beta}_1 = \{0.53, 0.37, 0.46, 0.57, ...\}$$

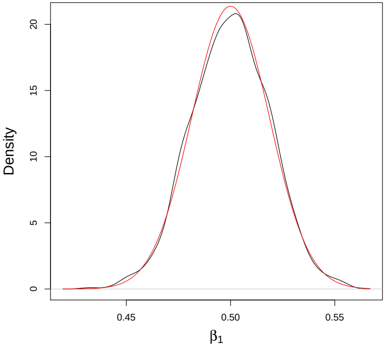
- Repeat this 1000 times and draw the density of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Properties of the OLS estimator - Normality



- This is the density of $\hat{\beta}_0$ estimated using 1000 samples, each of 100 observations, from our population.
- The red density is that of a Normal distribution with the same mean and variance. Note $\mathbb{E}[\hat{\beta}_0] \approx 10$.

Properties of the LRM estimators - Normality



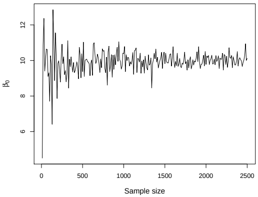
- This is the density of $\hat{\beta}_1$ estimated using 1000 samples, each of 100 observations, from our population.
- The red density is that of a Normal distribution with the same mean and variance. Note $\mathbb{E}[\hat{\beta}_1] \approx \frac{1}{2}$.

Properties of the OLS estimator

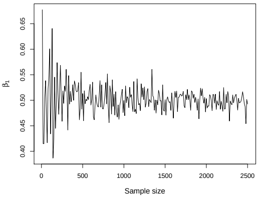
- The remarkable thing about the CLT is that it allows us to understand what would happen under repeated sampling just from the sample itself.
- That is, we can use it to learn about other (hypothetical) samples from our single sample. We do not need to take multiple samples to do this.

Properties of the OLS estimator - convergence

The Estimate of β_0



The Estimate of β_1



- The convergence towards their true values $(10, \frac{1}{2})$ is evident as is the reduction in their variability.
- But the convergence is not smooth which is why it is only "convergence in probability".

Statistical inference in the LRM

Recap

- A LRM coefficient $\hat{\beta}$ has a sampling distribution which, like that of an average,
  - has an expected value equal to its population value: $\mathbb{E}[\hat{\beta}] = \beta$.
  - has a standard error which we approximate/estimate by $\hat{SE}(\beta)$
  - has a sampling distribution which is asymptotically Normal.
- The normalised coefficient

$$t = \frac{(\hat{\beta} - \beta)}{\hat{SE}(\beta)}$$

has an asymptotically Standard Normal distribution: $N(0, 1)$.

Statistical inference in the LRM

Statistical Inference

Confidence intervals.

Testing hypotheses about a single regression parameter.

Measures of regression fit.

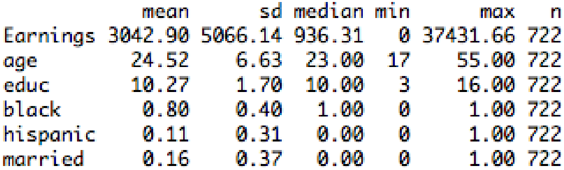
Testing hypotheses about several regression parameters.

Statistical inference in the LRM

- We will use some data from the National Supported Work (NSW) program.
- The NSW experiment was a temporary employment program in the US designed to help disadvantaged workers lacking basic job skills move into the labour market by giving them work experience and counselling.
- In what follows we will look at the relationship between an individual's earnings and their characteristics.
- We are interested in correlates of/determinants of earnings among the relevant population (low skill, low education males in the US).

Statistical inference in the LRM

- Our sample consists of 722 men from the experiment.
- The study collected data these men's age, race, schooling, marital status and their earnings.



```
            mean      sd median min      max   n
Earnings 3042.90 5066.14 936.31   0 37431.66 722
age        24.52    6.63  23.00  17    55.00 722
educ       10.27    1.70  10.00   3    16.00 722
black       0.80    0.40   1.00   0     1.00 722
hispanic    0.11    0.31   0.00   0     1.00 722
married     0.16    0.37   0.00   0     1.00 722
```

Statistical inference in the LRM

- We will specify the multivariate LRM as follows

$$\mathbb{E}\left[Y|X_1,...,X_4\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \left(X_2\right)^2 + \beta_4 X_3 + \beta_5 \left(X_3 X_4\right)$$

- Where

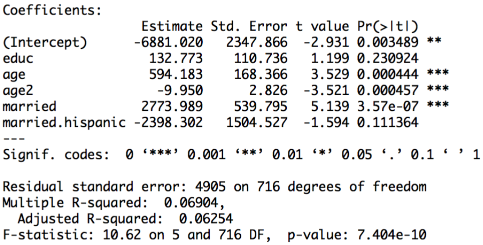
Variable	Characteristic
Y	Earnings ($\texttt{Earnings} \in \mathbb{R}_+$)
X_1	Education ($\texttt{educ} \in \mathbb{N}$)
X_2	Age ($\texttt{age} \in \mathbb{N}$)
X_3	Marital status ($\texttt{married} \in \{0,1\}$)
X_4	Hispanic American ($\texttt{hispanic} \in \{0,1\}$)

- Note that $\left(X_2\right)^2$ and $\left(X_3 X_4\right)$ are new variables which we made from the data.

Interpreting Regressions

$$\mathbb{E}\left[Y|X_1,...,X_4\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \left(X_2\right)^2 + \beta_4 X_3 + \beta_5 \left(X_3 X_4\right)$$

- The estimated regression



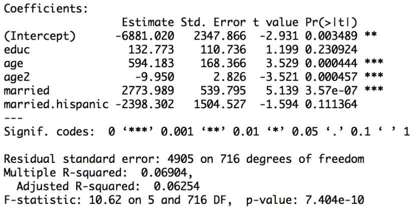
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -6881.020   2347.866  -2.931 0.003489 **
educ               132.773    110.736   1.199 0.230924
age                594.183    168.366   3.529 0.000444 ***
age2                -9.950      2.826  -3.521 0.000457 ***
married           2773.989    539.795   5.139 3.57e-07 ***
married.hispanic -2398.302   1504.527  -1.594 0.111364
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4905 on 716 degrees of freedom
Multiple R-squared:  0.06904,
  Adjusted R-squared:  0.06254
F-statistic: 10.62 on 5 and 716 DF,  p-value: 7.404e-10
```

- This is from R but it very typical of the kind of stuff which spews out of any statistical software.

Interpreting Regressions

$$\mathbb{E}\left[Y|X_1, ..., X_4\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \left(X_2\right)^2 + \beta_4 X_3 + \beta_5 \left(X_3 X_4\right)$$

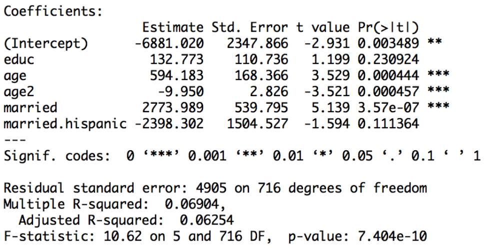


```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -6881.020   2347.866  -2.931 0.003489 **
educ               132.773    110.736   1.199 0.230924
age                594.183    168.366   3.529 0.000444 ***
age2                -9.950      2.826  -3.521 0.000457 ***
married           2773.989    539.795   5.139 3.57e-07 ***
married.hispanic -2398.302   1504.527  -1.594 0.111364
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4905 on 716 degrees of freedom
Multiple R-squared:  0.06904,
  Adjusted R-squared:  0.06254
F-statistic: 10.62 on 5 and 716 DF,  p-value: 7.404e-10
```

$-6881.020 = \hat{\beta}_0$ the sample estimate of the population coef. β_0.
$132.773 = \hat{\beta}_1$ the sample estimate of the population coef. β_1.
$594.183 = \hat{\beta}_2$ the sample estimate of the population coef. β_2.
$-9.950 = \hat{\beta}_3$ the sample estimate of the population coef. β_3.
$2773.989 = \hat{\beta}_4$ the sample estimate of the population coef. β_4.
$-2398.302 = \hat{\beta}_5$ the sample estimate of the population coef. β_5.

Statistical inference in the LRM



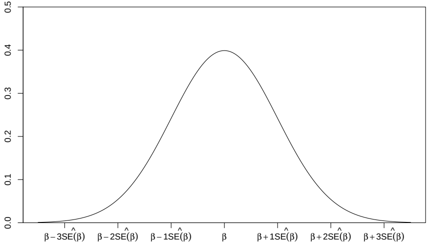
```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -6881.020   2347.866  -2.931 0.003489 **
educ                132.773    110.736   1.199 0.230924
age                 594.183    168.366   3.529 0.000444 ***
age2                 -9.950      2.826  -3.521 0.000457 ***
married            2773.989    539.795   5.139 3.57e-07 ***
married.hispanic -2398.302   1504.527  -1.594 0.111364
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4905 on 716 degrees of freedom
Multiple R-squared:  0.06904,
  Adjusted R-squared:  0.06254
F-statistic: 10.62 on 5 and 716 DF,  p-value: 7.404e-10
```

- The regression coefficients in the output tells us precisely about the association between Earnings and these other personal characteristics in the sample.
- But if we had had a different sample of men these numbers would surely have come out differently, but how different?
- What can we infer about the population values of the cofficients?

Statistical Inference

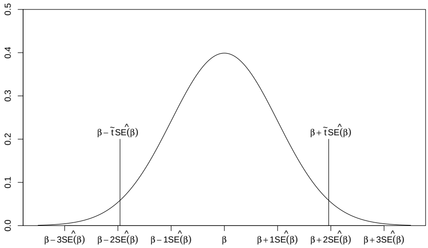
Confidence intervals for single regression parameter



- This shows the asymptotic sampling distribution of a LRM coefficient $\hat{\beta}$ for a sample of size n.
- Its mean $\mathbb{E}[\hat{\beta}]$ is the population parameter β with a standard error estimated by $SE\hat{}(\beta)$.

Statistical Inference

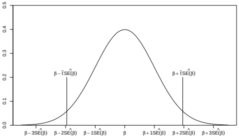
Confidence intervals on a regression parameter



- The probability that the value of $\hat{\beta}$ in repeated samples would lie within $\tilde{t}$ standard errors of its expected value (β) can be found from probability tables.

Statistical Inference

Confidence intervals on a regression parameter



$$Prob\left[\beta - \tilde{t}\widehat{SE}(\beta) \leq \hat{\beta} \leq \beta + \tilde{t}\widehat{SE}(\beta)\right] = c$$

- Note that this probability concerns the behaviour of the estimate $\hat{\beta}$ in repeated samples, not the behaviour of the population parameter β. To make statements about the population parameter β we need to rearrange.

Statistical Inference

Confidence intervals for single regression parameter

- Take the inequality

$$\beta - \tilde{t}\hat{SE}(\beta) \leq \hat{\beta} \leq \beta + \tilde{t}\hat{SE}(\beta)$$

- Take away β from all three terms

$$-\tilde{t}\hat{SE}(\beta) \leq \hat{\beta} - \beta \leq \tilde{t}\hat{SE}(\beta)$$

- Take away $\hat{\beta}$ from all three terms

$$-\hat{\beta} - \tilde{t}\hat{SE}(\beta) \leq -\beta \leq -\hat{\beta} + \tilde{t}\hat{SE}(\beta)$$

Statistical Inference

Confidence intervals for single regression parameter

- Multiply by minus one, (and reverse the inequalities)

$$\hat{\beta} + \tilde{t} \hat{SE}(\beta) \geq \beta \geq \hat{\beta} - \tilde{t} \hat{SE}(\beta)$$

- And finally re-order it smallest to biggest:

$$\hat{\beta} - \tilde{t} \hat{SE}(\beta) \leq \beta \leq \hat{\beta} + \tilde{t} \hat{SE}(\beta)$$

- If we substitute this back into the original probability we get

$$Prob\left[\hat{\beta} - \tilde{t} \hat{SE}(\beta) \leq \beta \leq \hat{\beta} + \tilde{t} \hat{SE}(\beta)\right] = c$$

Statistical Inference

Confidence intervals for single regression parameter

- The interpretation of a confidence interval is subtle. You should discuss it in tutorials. My preferred interpretation is ...

  "Were this procedure to be repeated on multiple samples, the calculated confidence interval (which would differ for each sample) would encompass the true population parameter $100c\%$ *of the time."*

- It's ungainly but it emphasises that estimates vary in repeated samples, and since the confidence interval is itself an estimated quantity it too would vary.

Statistical Inference

Confidence intervals for single regression parameter

- In the case of the coefficient on `married` in our regression

$$\hat{\beta} = 2773.989$$
$$\hat{SE}(\beta) = 539.795$$

- The 90% confidence interval is

$$\hat{\beta} \pm 1.645 \hat{SE}(\beta) = [1886.026,\ 3661.952]$$

- The 95% confidence interval is

$$\hat{\beta} \pm 1.96 \hat{SE}(\beta) = [1715.991,\ 3831.987]$$

- The 99% confidence interval is

$$\hat{\beta} \pm 2.58 \hat{SE}(\beta) = [1381.318,\ 4166.660]$$

Statistical inference in the LRM

Confidence intervals on a regression parameter

Confidence intervals

The $100(1-\alpha)\%$ confidence interval:

$$\hat{\beta} \pm \tilde{t} \hat{SE}(\beta)$$

where

$$\tilde{t} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

the value of which we look up in statistical tables.

Statistical inference in the LRM

Testing hypotheses about a single regression parameter

- A hypothesis test is a procedure which uses evidence from a sample to determine whether or not a specific hypothesis about a population is true or false.
- In the current context we are interested in hypotheses about the population value of an LRM coefficient.
- Statistical hypothesis testing is very easy to do, but quite hard to understand.

Statistical inference in the LRM

Testing hypotheses about a single regression parameter

- The hypothesis of interest is called the null hypothesis (denoted H_0) and its negation is known as the alternative hypothesis (denoted H_1).
- When we test a hypothesis about the population there are two sorts of errors we can make.
  - Type 1 errors: rejecting the null hypothesis when in fact it is true
  - Type 2 errors: failing to reject the null when in fact it is false.

Statistical inference in the LRM

Testing hypotheses about a single regression parameter

- When we carry out a hypothesis test we use a pre-specified rejection probability for the null when the null is true - i.e. the Type 1 error.
- For example, a significance level of 5% sets the probability of the Type 1 error (falsely rejecting the null) at 1 in 20.
- Sometimes we might say that we would “reject the null with 95% confidence” (i.e. only a 5% chance of doing so erroneously).

Statistical Inference

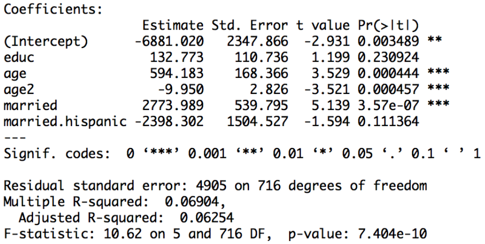
Testing hypotheses about a single regression parameter

Key elements:

1. State the null and alternative hypotheses.
2. Calculate the test statistic.
3. State its distribution under the null hypothesis.
4. State the decision rule and the decision.

Statistical Inference

Testing hypotheses about a single regression parameter



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -6881.020   2347.866  -2.931 0.003489 **
educ               132.773    110.736   1.199 0.230924
age                594.183    168.366   3.529 0.000444 ***
age2                -9.950      2.826  -3.521 0.000457 ***
married           2773.989    539.795   5.139 3.57e-07 ***
married.hispanic -2398.302   1504.527  -1.594 0.111364
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4905 on 716 degrees of freedom
Multiple R-squared:  0.06904,
  Adjusted R-squared:  0.06254
F-statistic: 10.62 on 5 and 716 DF,  p-value: 7.404e-10
```

- We are concerned that education and earnings are not in fact related in the population.
- Perhaps our estimate that $\hat{\beta} = 132.773$ was merely an artefact thrown up by chance in this particular sample.

Statistical Inference

Testing hypotheses about a single regression parameter

1. State the hull and alternative hypothesis:

$$H_0 : \beta_{\text{educ}} = 0 \qquad H_1 : \beta_{\text{educ}} \neq 0$$

2. Calculate the test statistics

$$t = \frac{\hat{\beta}_{\text{educ}} - 0}{SE\left(\hat{\beta}_{\text{educ}}\right)} = \frac{132.773 - 0}{110.736} = 1.199$$

3. State its distribution under the null hypothesis

Under the null $t \sim N(0, 1)$

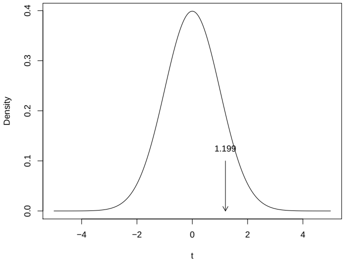
4. State the decision rule and the decision:

Reject H_0 with 5% significance or 95% confidence if $|t| > 1.96$

Decision: We cannot reject the null.

Statistical Inference

Testing hypotheses about a single regression parameter



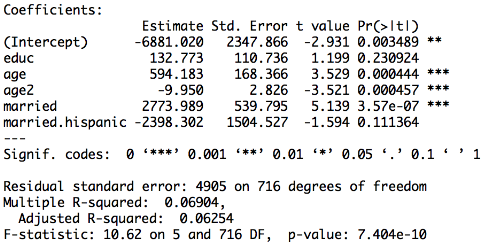
- If the null hypothesis were true ("under the null") and $\beta = 0$ then

$$t = \frac{\hat{\beta}_{\text{educ}} - 0}{SE\left(\hat{\beta}_{\text{educ}}\right)} \sim N(0,1)$$

- The observed value of t (1.199) is quite likely to be observed under the null.
- We therefore conclude that the evidence against the null is not too strong.

Statistical Inference

Testing hypotheses about a single regression parameter



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -6881.020   2347.866  -2.931 0.003489 **
educ               132.773    110.736   1.199 0.230924
age                594.183    168.366   3.529 0.000444 ***
age2                -9.950      2.826  -3.521 0.000457 ***
married           2773.989    539.795   5.139 3.57e-07 ***
married.hispanic -2398.302   1504.527  -1.594 0.111364
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4905 on 716 degrees of freedom
Multiple R-squared:  0.06904,
  Adjusted R-squared:  0.06254
F-statistic: 10.62 on 5 and 716 DF,  p-value: 7.404e-10
```

- The regression says that earnings rise with age but at a declining rate and indeed fall after the age of about 30.
- We want to know whether this is relationship holds in the population.

Statistical Inference

Testing hypotheses about a single regression parameter

1. State the hull and alternative hypothesis:

$$H_0 : \beta_{\text{age}^2} = 0 \qquad H_1 : \beta_{\text{age}^2} \neq 0$$

2. Calculate the test statistics

$$t = \frac{\hat{\beta_{\text{age}^2}} - 0}{SE\left(\hat{\beta}_{\text{age}^2}\right)} = \frac{-9.950 - 0}{2.826} = -3.521$$

3. State its distribution under the null hypothesis

$$\text{Under the null } t \sim N(0, 1)$$

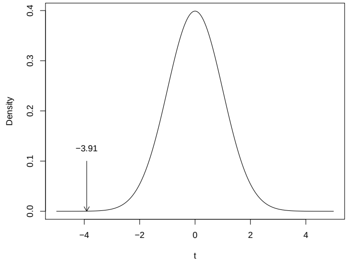
4. State the decision rule and the decision:

Reject H_0 with 5% significance or 95% confidence if $|t| > 1.96$

Decision: We reject the null.

Statistical Inference

The Object of Interest



- Under the null that $\beta_{\text{age}^2} = 0$

$$t = \frac{\hat{\beta}_{\text{age}^2} - 0}{SE\left(\hat{\beta}_{\text{age}^2}\right)} \sim N(0,1)$$

- The observed value of t (-3.521) is indicated. It is very unlikely to be observed under the null - you'd see values like this, this far from the null, very rarely.
- The null is probably not true.

Statistical Inference

Testing hypotheses about a single regression parameter

- Note that standard “significance tests” are normally tests of a null of “no effect”, i.e. $\beta = 0$.
- These are the kinds of t-values routinely reported by statistical software and in published articles.
- But you might be interested in tests of other hypothesis. For example, if the regression is in logs then you might be interest in whether the elasticity is one: $\beta = 1$.
- The procedure is exactly the same, the only difference is the value of β you assume under the null when you construct t.

Statistical Inference

Testing hypotheses about a single regression parameter

The alternative hypothesis we have looked at is

$$H_1 : \beta \neq 0$$

is sometimes called a "two-sided" alternative. Sometimes we are interested in testing for one-sided alternatives such as

$$H_0 : \beta = 0 \quad H_1 : \beta > 0$$

or

$$H_0 : \beta = 0 \quad H_1 : \beta < 0$$

Directional tests are based on a signed t-statistic.

Statistical Inference

Testing hypotheses about a single regression parameter

- Tests of

$$H_0 : \beta = 0 \quad H_1 : \beta > 0$$

  are based on the statistic

$$t = \frac{\hat{\beta} - 0}{\hat{SE}(\beta)}$$

- We reject the null at 95% if

$$t > 1.645$$

  (the 5% critical value from the upper tail of the Std. Normal).
- Negative values of t are not taken as evidence against the null as estimates of $\hat{\beta}$ less than 0 do not point towards H_1.

Statistical Inference

Testing hypotheses about a single regression parameter

- Tests of

$$H_0 : \beta = 0 \quad H_1 : \beta < 0$$

are based on the statistic

$$t = \frac{0 - \hat{\beta}}{\hat{SE}(\beta)}$$

- We reject the null at 95% if

$$t < -1.645$$

(the lower tail 5% critical value).

- Positive values of t are not taken as evidence against the null as estimates of $\hat{\beta}$ greater than 0 do not point towards H_1.

Statistical Inference

Testing hypotheses about a single regression parameter

- Since the critical values of the test statistics are taken from the single tail of the standard normal they are smaller than for two-sided tests.
- There seems to be an ambiguity. Should we use the two-sided critical value 1.96 or the one- sided critical value 1.645? The answer is that we should use one-sided tests and critical values only when the parameter space is known to satisfy a one-sided restriction. This is when the one-sided test makes sense.
- If such a restriction is not known *a priori*, then imposing this restriction to form the test does not make sense.
- Two-sided tests are generally appropriate.

Statistical Inference

Testing hypotheses about a single regression parameter

- Testing requires a pre-selected choice of significance level (probability of a Type 1 error), yet there is no objective scientific basis for this.
- Common practice is to use 5% for the significance (or 95% for confidence level).
- The informal reasoning behind the choice of a 5% critical value is to ensure that Type I errors (false positives) should be relatively unlikely — that the decision "Reject H_0" should have scientific strength — yet the test retains power against reasonable alternatives.

Statistical Inference

Testing hypotheses about a single regression parameter - P values

- The decision “Reject H_0” means that the evidence is inconsistent with the null hypothesis, in the sense that it is relatively unlikely (say less than 1 in 20) that data generated by the null hypothesis would yield the observed test result.
- In contrast, the decision “Accept H_0” is not a strong statement. It does not mean that the evidence supports H_0, only that there is insufficient evidence to reject H_0.
- Because of this, it is more accurate to use the formulation “Do not Reject H_0” instead of “Accept H_0”.

Statistical Inference

Testing hypotheses about a single regression parameter - P values

- When we look at some regression results we often find that some t-stats are bigger than other.
- Some may just limp over the two-tail critical value whereas others may just fall beneath it.
- For example, if one coefficient has a t-stat of 1.9, which is less than the critical value and another had a t-stat of 2.0, the latter would be hailed as “significant” whereas the former written off as “insignificant”.
- Yet their t’s are close.

Statistical Inference

Testing hypotheses about a single regression parameter - P values

- Should we really be making a different decision if the t-statistic is 1.9 rather than 2.0? The difference in values is small, shouldn't the difference in the decision be also small?
- This suggest that simply putting variables into somewhat slightly arbitrary "bins" based on a rigid 1.96 cut-off could be unsatisfactory as it does not really summarise the evidence.
- Instead, the magnitude of the statistic may suggest a "degree of evidence" against H_0

Statistical Inference

Testing hypotheses about a single regression parameter - P values

- How can we take this into account? The answer is to report what is known as the asymptotic p-value.
- For a two-sided test it is given by

  $$p = 2(1 - \Phi(t))$$

  or equivalently

  $$p = 2(\Phi(-|t|))$$

- In order to understand a p-value, you must first understand what the null hypothesis is.
- Standard statistical software reports p-values for a null of "no effect": $H_0 : \beta = 0$

Statistical Inference

Testing hypotheses about a single regression parameter - P values

- The p-value evaluates how well the sample data supports the null of “no effect”.
- High p-values: your data are likely under the null. (You are very likely to see the effect observed in your sample data if the null hypothesis is true)
- Low p-values: suggests that your sample provides enough evidence that you can reject the null hypothesis.(You are very unlikely to see the effect observed in your sample data if the null hypothesis is true).

Statistical Inference

Testing hypotheses about a single regression parameter - P values

- It is instructive to interpret the p-value as the marginal significance level: the largest value of the significance level or which the test rejects the null hypothesis.
- An important caveat is that the p-value should not be interpreted as the probability that either hypothesis is true.
- For example, a common mis-interpretation is that it is the probability “that the null hypothesis is false.” This is incorrect. Rather, it is a measure of the strength of information against the null hypothesis.

Statistical Inference

Testing hypotheses about a single regression parameter - P values

- Returning to our empirical example, we can see that p-values on `age` and `age2` are effectively zero (they are not actually zero, just very very close).
- When presented with such evidence we can say that we “strongly reject” the null hypothesis, that the test is “highly significant”, or that “the test rejects at any conventional critical value”.

Statistical Inference

Testing hypotheses about a single regression parameter - P values

- Recall the odd coefficient on `married.hispanic`? The p-value on a null if no effect is 0.111364.
- There is quite a high probability of observing the coefficient of -2398.302 in a sample even when the value in the population was zero.
- Indicates that the relationship is statistically insignificant at most conventional test levels.
- We would fail to reject the null of no effect at 95% (1.96).
- At a significance level of 11% (confidence of 89%) we would however just be on the cusp of not rejecting H_0.

Statistical Inference

Testing hypotheses about a single regression parameter

- Standard statistical software normally report p-values and t-vales for null of "no effect".
- Informally - although highly practically - assessing variable-by-variable statistical significance against this null amounts to little more than have a quick squizz down the list of t-stats or p-values and noting
  - which t's are bigger than/less than 1.96 and/or
  - which p-values are bigger/less than 0.05.
- Of course in examinations it's essential to set out the test formally.
- You also need to be able to handle statistical tables (in case the Examiners ask for the 91% confidence interval or to carry out a test at 13.5% significant etc).

$$Y = \beta_0 + \beta_1 X + u$$

- Since the linear model describes the data as the sum of a linear function of the regressor X and a residual which is uncorrelated with it we can write

$$Var\left(Y\right) = Var\left(\beta_0 + \beta_1 X\right) + Var\left(u\right)$$

- This decomposes the variability in the data into a part which is described by the variability captured by the model, and a part which is not so captured.
- This mirrors the similar decomposition with the CEF.

Statistical Inference

Assessing the fit - R-squared

- Recall that the variance of the data is

$$var(Y) = \frac{\sum_i (Y_i - \bar{Y})^2}{n-1}$$

- The numerator is called the “Total Sum of Squares” (TSS).
- Thus

$$TSS \propto Var(Y)$$

- Similarly $Var(\beta_0 + \beta_1 X)$ is proportional to what we call the “Explained Sum of Squares” (ESS) and $Var(u)$ is proportional to the “Sum of Squared Residuals” or (SSR).

Statistical Inference

Assessing the fit - R-squared

- The factors of proportionality (the $n-1$ terms) cancel so we generally write the decomposition of the variance as

$$TSS = ESS + SSR$$

- This says that the overall variation in the data is the sum of the variation captured in the model and the variation in the residual.
- We can then quantify what proportion of the total variation is accounted for by the model:

$$\frac{ESS}{TSS}$$

  and what proportion is unexplained

$$\frac{SSR}{TSS}$$

Statistical Inference

Assessing the fit - R-squared

- The term

$$R^2 = \frac{ESS}{TSS}$$

  or equivalently

$$R^2 = 1 - \frac{SSR}{TSS}$$

  is the population R^2 or the coefficient of determination.

Statistical Inference

Assessing the fit - R-squared

- In the example we have (not reported in the regression output given)

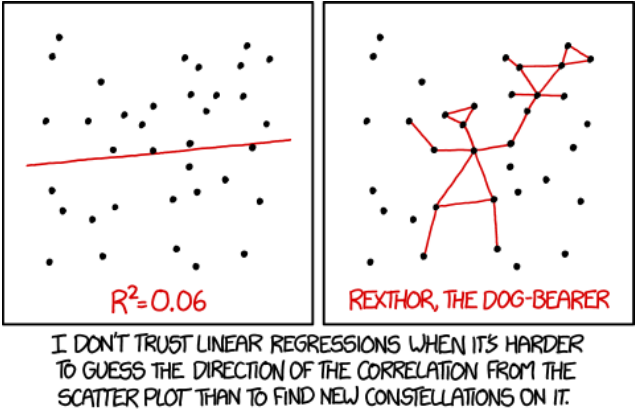
$$TSS = ESS + SSR$$
$$18505048001 = 1277606200 + 17227441800$$

and

$$R^2 = \frac{1277606200}{18505048001} = 1 - \frac{17227441800}{18505048001} = 0.06904$$

Statistical Inference

Assessing the fit - R-squared


R^2=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Source: xkcd

Statistical Inference

Assessing the fit - R-squared

- Introducing an additional regressor into the regression model generally reduces the SSR and hence increases the R^2 even if the regressor has negligible explanatory power.
- Comparing models with different numbers of parameters on the basis of R^2 is not sensible - the model with most parameters cannot lose.
- But Occam's Razor would suggest that the best model is one which can fit the best whilst using the fewest number of parameters in order to do it.

- The "adjusted R^2" denoted $\bar{R}^2$ corrects this problem by "penalising" you for including another regressor. It does not necessarily increase when you add another regressor.

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-K-1}\right)\frac{SSR}{TSS}$$

Where n is the number of observations and K is the number of parameters (not including the intercept). Note that

$$\left(\frac{n-1}{n-K-1}\right) > 1 \Rightarrow R^2 > \bar{R}^2$$

Statistical Inference

Assessing the fit - R-squared

- As we add regressors SSR will naturally go down tending to push $\bar{R}^2$ up.
- But
$$\left(\frac{n-1}{n-K-1}\right)$$
will go up, pushing $\bar{R}^2$ in the opposite direction - possibly enough to make the value of $\bar{R}^2$ negative.
- In our example
$$\bar{R}^2 = 1 - \left(\frac{722-1}{722-5-1}\right)\frac{17227441800}{18505048001} = 0.06254$$

Statistical Inference

Assessing the fit - R-squared

- Some points to note
    - an R^2 or $\bar{R}^2$ close one means that the explanatory variables are good at fitting the data Y; values close to zero mean that they are not.
    - high values of R^2 or $\bar{R}^2$ do not tell you necessarily that the model will be good at predicting (extrapolating) "out of sample".
    - high values of R^2or $\bar{R}^2$ do not tell you anything about causality.
    - while R^2 provides an estimate of the strength of the relationship between your model and the response variable, it does not provide a formal hypothesis test for this relationship.

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- We may wish to consider joint hypotheses about several population coefficients.
- Individual t-tests are not suitable for this because that distorts the size of the joint test (i.e. the actual rejection rate under the null).
- Instead we use a test statistic that takes into account the joint distribution of the estimators: the F-test.

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- The F distribution is used to describe random variables of certain types - just as the Normal is used to describe random variables of certain types (e.g. standardised means).
- Just as the Normal is a curve whose shape is determined by two parameters the F is a curve governed by two parameters

$$F(d_1, d_2)$$

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- The most common hypothesis which we examine as a matter of course in applied empirical work is the hypothesis that, in the population, “nothing matters”.
- This is the joint hypothesis that all of the population regression slope parameters are jointly zero

$$H_0 : \beta_1 = \beta_2 = ...\beta_K = 0$$

- This a test of the joint significance of all of the variables in the regression.
- The two parameters in the F distribution are (in this case)

$$d_1 = K$$
$$d_2 = n - K - 1$$

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- Typically, if you don't have any significant p-values or t-stats for the individual coefficients in your model, the overall F-test won't be significant either.
- However, in a few cases, the tests could yield different results.
- For example, a significant overall F-test could determine that the coefficients are *jointly* not all equal to zero while the tests for individual coefficients could determine that all of them are *individually* equal to zero.
- This is particularly the case when the regressors are close to being co-linear.

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- Typically regression output reports the “omnibus F-test”: a test of the hypothesis that “nothing matters”.
- The unrestricted model is

$$Y = \beta_0 + \beta_1 X_1 + ...\beta_K X_K + \epsilon$$

- The restricted model is

$$Y = \beta_0 + \varepsilon$$

- There are K restrictions: $\beta_k = 0$ for $k = 1, ..., K$.

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- The F test is based on the question of whether or not the use of all of the regressors in the full model significantly (in the statistical sense) reduces the unexplained variation in the data compared to not using the regressors.
- It centres on the SSR (the measure of the variation in the data which the model cannot explain) and compares it between the restricted and the unrestricted models.

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- The SSR_{rest} is the SSR from the restricted model, and SSR_{unrest} is the SSR from the unrestricted model, then the F statistic is

$$F = \frac{\left[\frac{SSR_{restr} - SSR_{unrest}}{K}\right]}{\left[\frac{SSR_{unrest}}{n-K-1}\right]}$$

or

$$F = \left[\frac{n-K-1}{K}\right]\frac{SSR_{rest} - SSR_{unrest}}{SSR_{unrest}}$$

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- Alternatively you can use the fact that

$$TSS_{unrest} = SSR_{rest}$$

and then use

$$F = \frac{\left[\frac{TSS_{unrestr} - SSR_{unrest}}{K}\right]}{\left[\frac{SSR_{unrest}}{n-K-1}\right]}$$

or

$$F = \left[\frac{n-K-1}{K}\right]\frac{TSS_{unrest} - SSR_{unrest}}{SSR_{unrest}}$$

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- Null hypothesis: $H_0 : \beta_1 = ... = \beta_K = 0$
- Alternative hypothesis: $H_1 : \beta_k \neq 0$ for at least one regressor.
- Test statistic: $F = \left[\frac{n-K-1}{K}\right] \frac{TSS_u - SSR_u}{SSR_u} \sim F(K, n-K-1)$
- Critical value: $c_\alpha = F^{-1}(1-\alpha)$
- Decision rule: Reject H_0 if $F > c_\alpha$.

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- In our running example regression an F-statistic of "10.62 on 5 and 716 DEF", is reported.
- This is test between the model as estimated and the restricted model which has

  $$\beta_1 = \beta_2 = ...\beta_K = 0$$

  i.e. a model with just a constant.
- There are $K = 5$ restrictions (one for each regressor) and $n - K - 1 = 722 - 5 - 1 = 716$ thus the statistic has the parameters

  $$F(5, 716)$$

- This 95% critical value for this is 2.12.

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- You can compute the value of the F stat from the sums-of-squares

$$F = \left[\frac{n - K - 1}{K}\right] \frac{TSS_{unrest} - RSS_{unrest}}{RSS_{unrest}}$$

$$F = \left[\frac{722 - 5 - 1}{5}\right] \frac{18505048001 - 17227441800}{17227441800} = 10.62$$

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- There are a couple of additional conclusions you can draw from a significant overall F-test.
- In the intercept-only model, all of the fitted values equal the mean of the response variable. Therefore, if the p-value of the overall F-test is significant, your regression model predicts the response variable better than the mean of the response.

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- While R-squared provides an estimate of the strength of the relationship between your model and the response variable, it does not provide a formal hypothesis test for this relationship.
- The overall F-test determines whether this relationship is statistically significant. If the p-value for the overall F-test is less than your significance level, you can conclude that the R-squared value is significantly different from zero.

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- The idea that our explanatory variables have no statistically significant relationship with the dependent variable is clearly related to the R^2 which tell us how much of the overall variability in Y is explained by the LRM.
- In fact under some additional assumptions about the variance of the errors in the LRM (i.e. their variance is constant for all values of the X's) there is an exact relationship between the two.

Statistical Inference

Testing hypotheses about several regression parameters - the F test

- This is based on a comparison between the R^2 of the regression we estimated and the R^2 of a regression which doesn't have any explanatory variables at all.

$$F(K, n-K-1) = \left[\frac{n-K-1}{K}\right] \frac{R^2_{unrest} - R^2_{rest}}{1 - R^2_{unrest}}$$

K is the number of regressors in the unrestricted model.

Summary

Confidence intervals

The $100(1-\alpha)\%$ confidence interval:

$$t^* = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

$$\hat{\beta} \pm t^* \hat{SE}(\beta)$$

Summary

t tests

Test-statistic to test if the population coefficient β equals a certain value β_0 at the α significance level:

Null hypothesis: $H_0 : \beta = \beta_0$

Alternative hypothesis: $H_1 : \beta \neq \beta_0$

Test statistic: $t = \frac{\beta - \beta_0}{\hat{SE}(\beta)} \sim N(0,1)$

Critical value: $c_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$

Decision rule: Reject H_0 if $|t| > c_\alpha$.

Summary

p-values

Probability of observing a value as or more extreme than t, when the null hypothesis is true.

$$p = 2(1 - \Phi(t))$$

$$\begin{aligned} t = 2.575 \text{ gives } p &= 2(1 - \Phi(2.575)) \approx 0.01 \\ t = 1.96 \text{ gives } p &= 2(1 - \Phi(1.96)) \ \approx 0.05 \\ t = 1.645 \text{ gives } p &= 2(1 - \Phi(1.645)) \approx 0.10 \end{aligned}$$

Summary

Omnibus F test

The "null model" is $Y = \beta_0 + \varepsilon$
The "full model" is $Y = \beta_0 + \beta_1 X_1 + ...\beta_K X_K + \epsilon$

Null hypothesis: $H_0 : \beta_1 = ... = \beta_K = 0$

Alternative hypothesis: $H_1 : \beta_k \neq 0$ for at least one regressor.

Test statistic: $F = \frac{n-K-1}{K}\frac{TSS-SSR}{SSR} \sim F(K, n-K-1)$

Critical value: $c_\alpha = F^{-1}(1-\alpha)$

Decision rule: Reject H_0 if $F > c_\alpha$.

Summary

R^2 and $\bar{R}^2$

The “coefficient of determination”

$$R^2 = 1 - \frac{SSR}{TSS}$$

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right) \frac{SSR}{TSS}$$

Summary

- You need to know the definitions of and be able to construct
  - confidence intervals,
  - t tests,
  - p values,
  - R^2 and $\bar{R}^2$
  - F tests.
- You need to be able to do this on the basis of whatever information is presented to you (this requires you to understand the connections between them).
- Just as important (and possibly more so) you need to be able to show that you understanding the meaning and interpretation of all of these statistics.