London Business School

Session 9: Regression models; regression diagnostics

Kostis Christodoulou
London Business School

Contents

- Further regression examples
- Interaction variables
- Regression Diagnostics

London Business School

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

Data is not labeled in any way

SUPERVISED

Predict a category

Predict a number

CLASSIFICATION

«Divide the socks by color»

REGRESSION

«Divide the ties by length»

UNSUPERVISED

Divide by similarity

Find hidden dependencies

Identify sequences

CLUSTERING

«Split up similar clothing into stacks»

ASSOCIATION

«Find what clothes I often wear together»

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

https://vas3k.com/blog/machine_learning/

Import- Inspect- Clean- Explore Data

Import

- Use vroom() or read_csv()

Inspect

- Use skimr::skim()
- Understand each variable definition
- Note the units variables are measured in (do not confuse lbs. with Kg)
- Identify any data issues (missing values, incorrect entries, etc.)

Clean

- Use ***janitor::clean_names()***
- Decide what to do with missing values. Impute?
- Code variables correctly (e.g., numerical variables, factors, dates)
- Use ***tidyr::pivot_longer()*** and/or ***tidyr::pivot_wider()*** to rearrange the data set if it needs to be in tidy format

Explore

- Articulate hypotheses as to what may be happening – discuss with colleagues / experts
- EDA: correlation charts, histograms, scatter plots, etc

London Business School

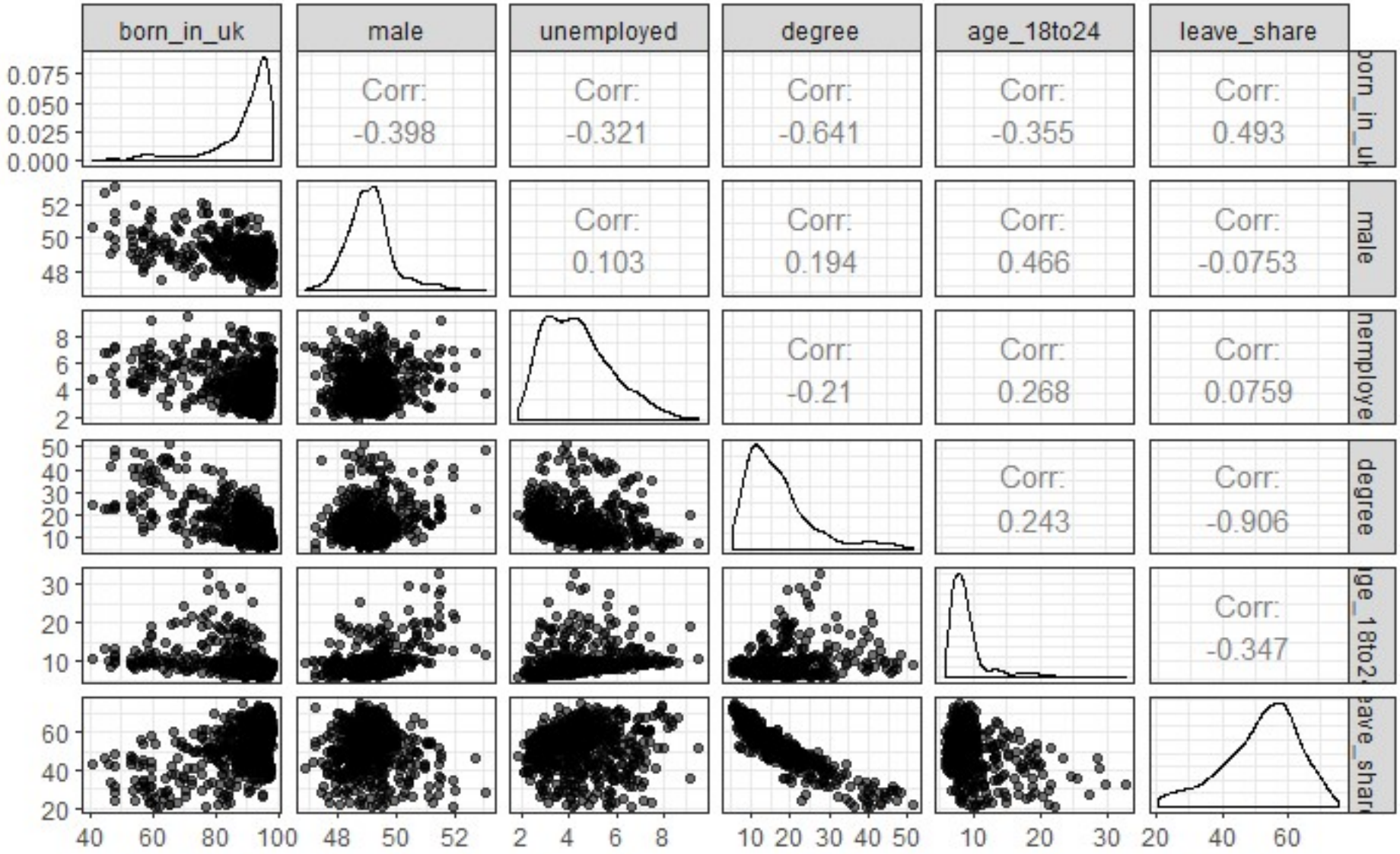
Brexit data

Filter

	Seat	con_2015	lab_2015	ld_2015	ukip_2015	leave_share	born_in_uk	male	unemployed	degree	age_18to24
1	Aldershot	50.592	18.333	8.824	17.867	57.89777	83.10464	49.89896	3.637000	13.870661	9.406093
2	Aldridge-Brownhills	52.050	22.369	3.367	19.624	67.79635	96.12207	48.92951	4.553607	9.974114	7.325850
3	Altrincham and Sale West	52.994	26.686	8.383	8.011	38.58780	90.48566	48.90621	3.039963	28.600135	6.437453
4	Amber Valley	43.979	34.781	2.975	15.887	65.29912	97.30437	49.21657	4.261173	9.336294	7.747801
5	Arundel and South Downs	60.788	11.197	7.192	14.438	49.70111	93.33793	48.00189	2.468100	18.775591	5.734730
6	Ashfield	22.418	41.022	14.828	21.409	70.47289	96.96214	49.17185	4.742731	6.085457	8.209863
7	Ashford	52.454	18.441	5.984	18.821	59.86195	90.50823	48.52222	3.687889	13.121731	7.815654
8	Ashton-under-Lyne	22.123	49.761	2.423	21.759	61.80980	90.72875	49.17554	5.108282	7.899545	8.937492
9	Aylesbury	50.674	15.141	10.619	19.713	51.78791	86.95974	49.51262	3.390869	17.798940	7.561073
10	Banbury	53.008	21.297	5.930	13.877	50.34780	88.82538	49.45814	2.932093	16.700324	7.606336
11	Barking	16.308	57.680	1.306	22.197	59.97488	66.39278	48.90159	7.364821	14.440739	9.631029
12	Barnsley Central	15.003	55.733	2.106	21.720	68.18813	95.60282	49.35016	5.337255	8.134895	8.608862
13	Barnsley East	14.596	54.726	3.160	23.483	70.98499	97.13493	48.95062	5.620404	6.520969	8.345313
14	Barrow and Furness	40.497	42.334	2.701	11.716	57.27871	96.96598	49.37181	4.202689	10.665089	7.775876
15	Basildon and Billericay	52.682	23.673	3.802	19.843	67.13588	92.50616	48.17421	4.613724	10.881236	7.790427
16	Basingstoke	48.551	27.707	7.384	15.619	53.59878	87.02730	49.44535	3.631748	16.519574	8.045668
17	Bassetlaw	30.680	48.621	2.700	15.957	68.32276	95.24218	49.49041	4.016112	9.035628	7.798050
18	Bath	37.808	13.179	29.682	6.195	31.72455	86.22987	48.99335	2.777410	28.903857	18.675655
19	Batley and Spen	31.239	43.238	4.747	17.988	59.62840	90.17044	49.35820	4.939431	10.560815	8.697949
20	Battersea	52.380	36.825	4.391	3.108	22.04674	65.48370	48.89278	3.849259	51.098323	9.022669

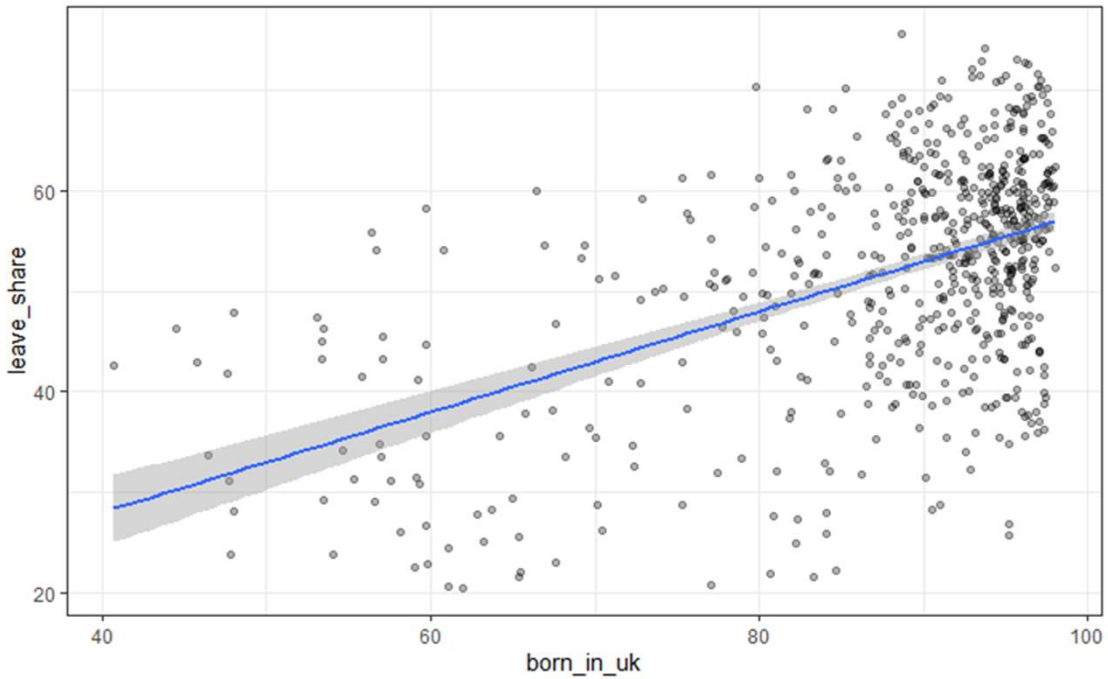
London Business School

Brexit correlations



London Business School

Brexit model 1

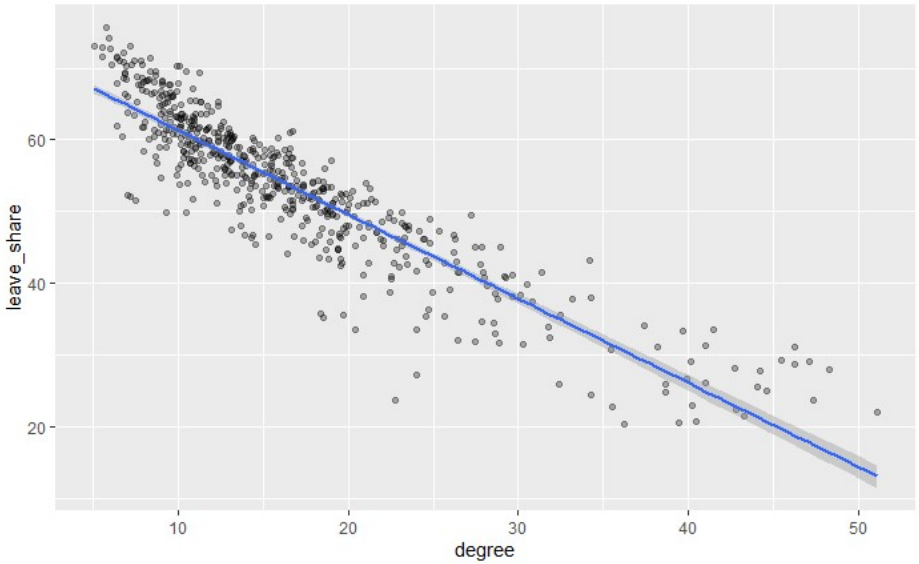


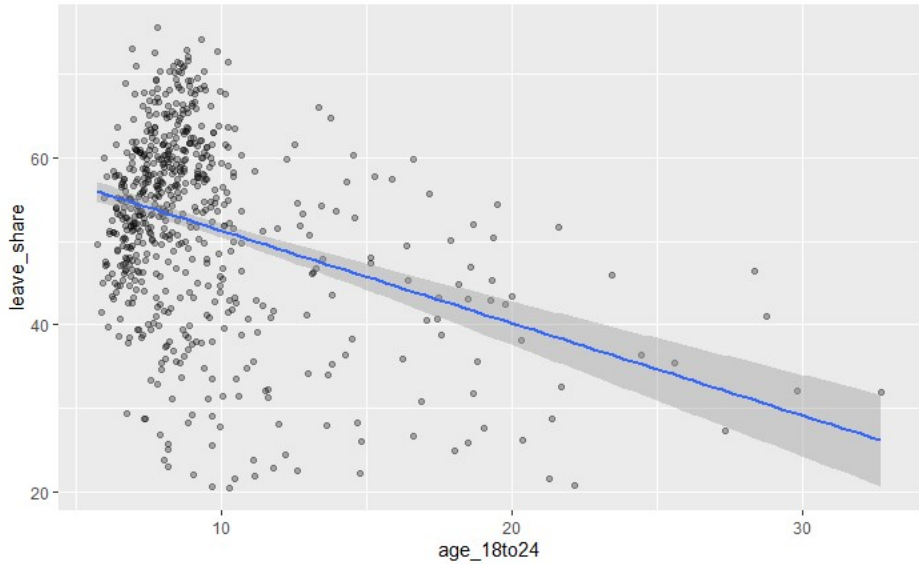
term <chr>	estimate <dbl>	std_error <dbl>	statistic <dbl>	p_value <dbl>	lower_ci <dbl>	upper_ci <dbl>
intercept	7.977	3.121	2.556	0.011	1.848	14.106
born_in_uk	0.500	0.035	14.239	0.000	0.431	0.569

r_squared <dbl>	adj_r_squared <dbl>	mse <dbl>	rmse <dbl>	sigma <dbl>	statistic <dbl>	p_value <dbl>	df <dbl>
0.243	0.242	98.82176	9.940913	9.957	202.752	0	2

London Business School

Brexit model 2



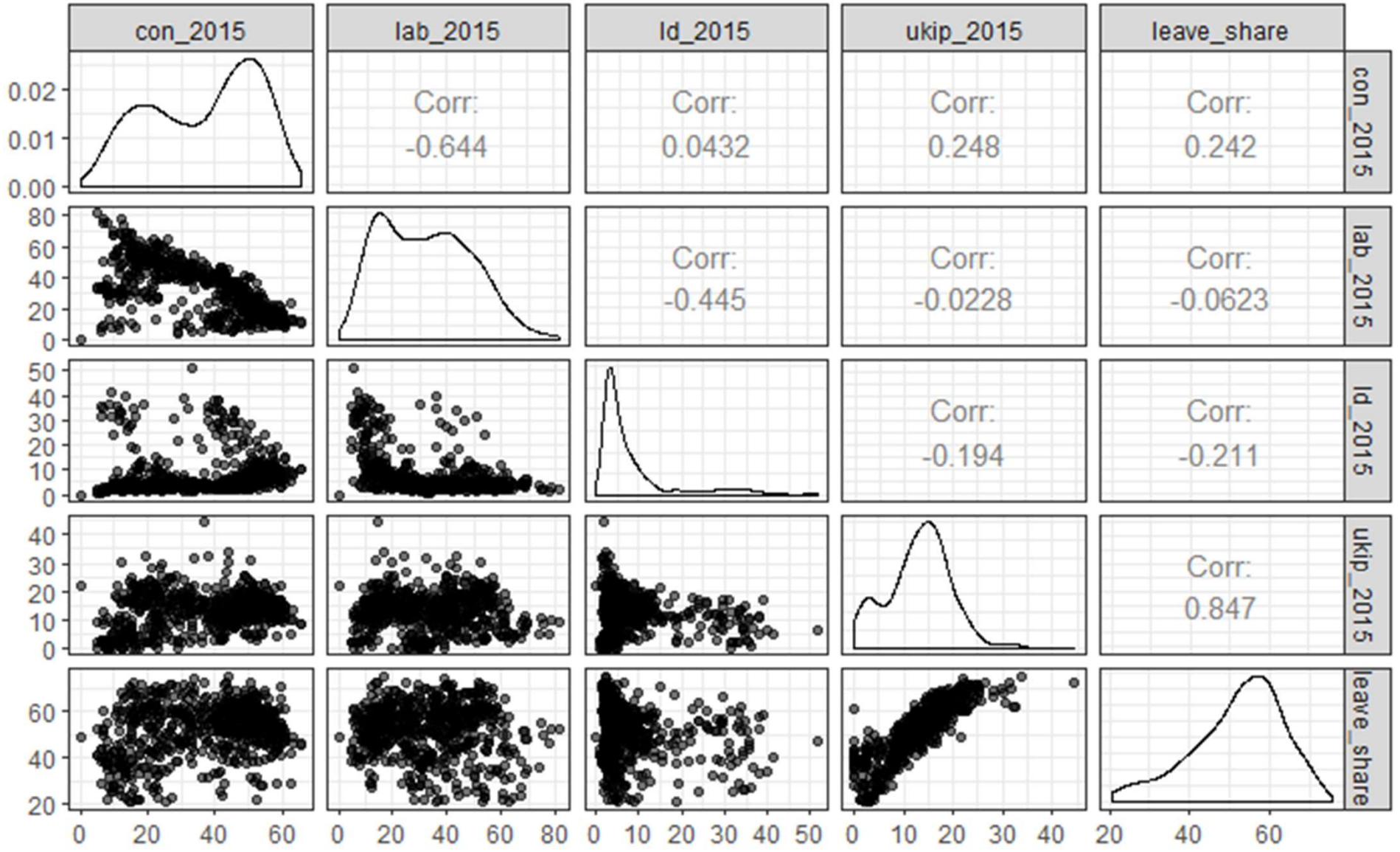


term <chr>	estimate <dbl>	std_error <dbl>	statistic <dbl>	p_value <dbl>	lower_ci <dbl>	upper_ci <dbl>
intercept	76.510	0.558	137.123	0	75.414	77.606
degree	-1.125	0.022	-50.789	0	-1.169	-1.082
age_18to24	-0.456	0.052	-8.847	0	-0.557	-0.355

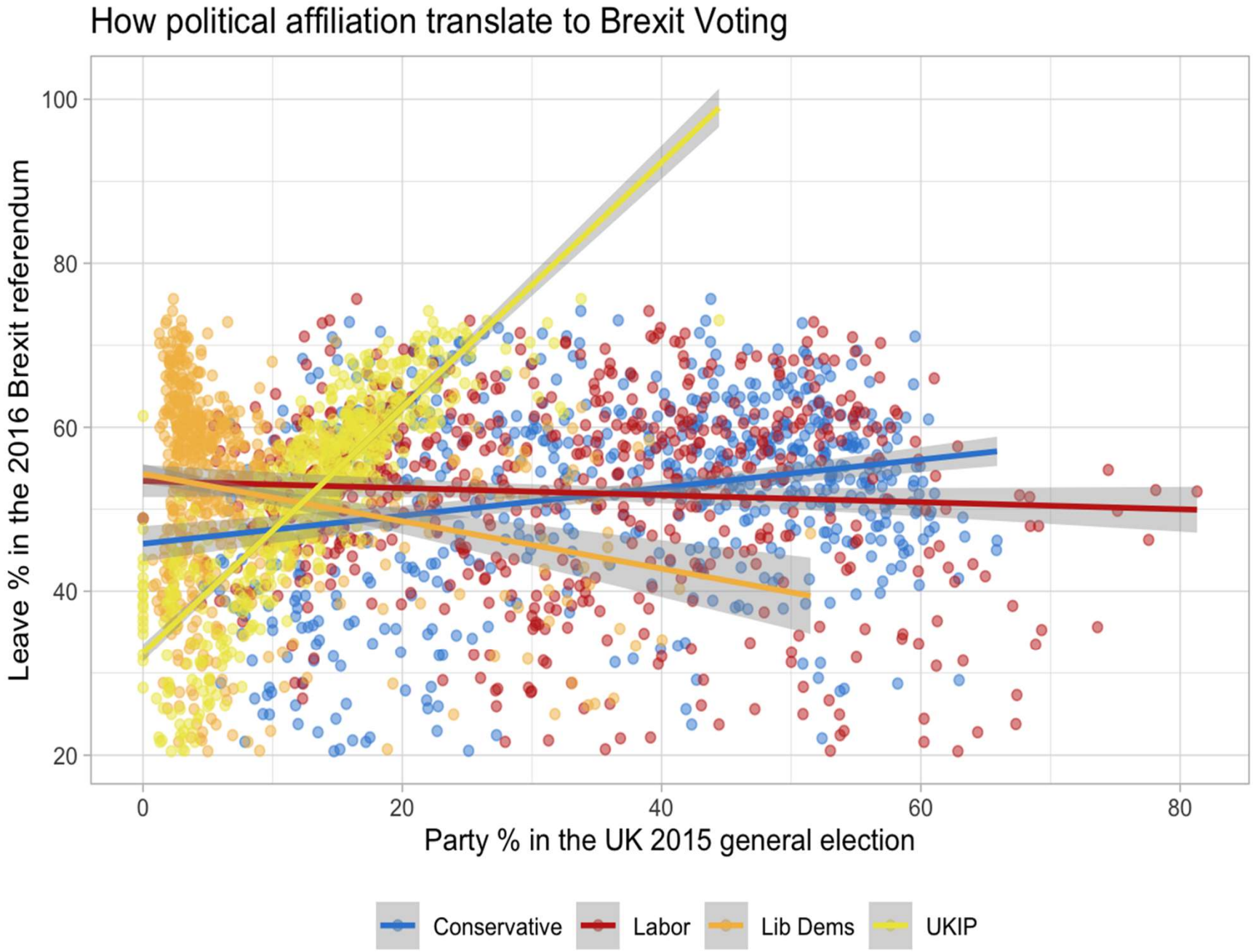
r_squared <dbl>	adj_r_squared <dbl>	mse <dbl>	rmse <dbl>	sigma <dbl>	statistic <dbl>	p_value <dbl>	df <dbl>
0.843	0.842	18.34363	4.282946	4.294	1527.929	0	3

London Business School

Brexit and party affiliation

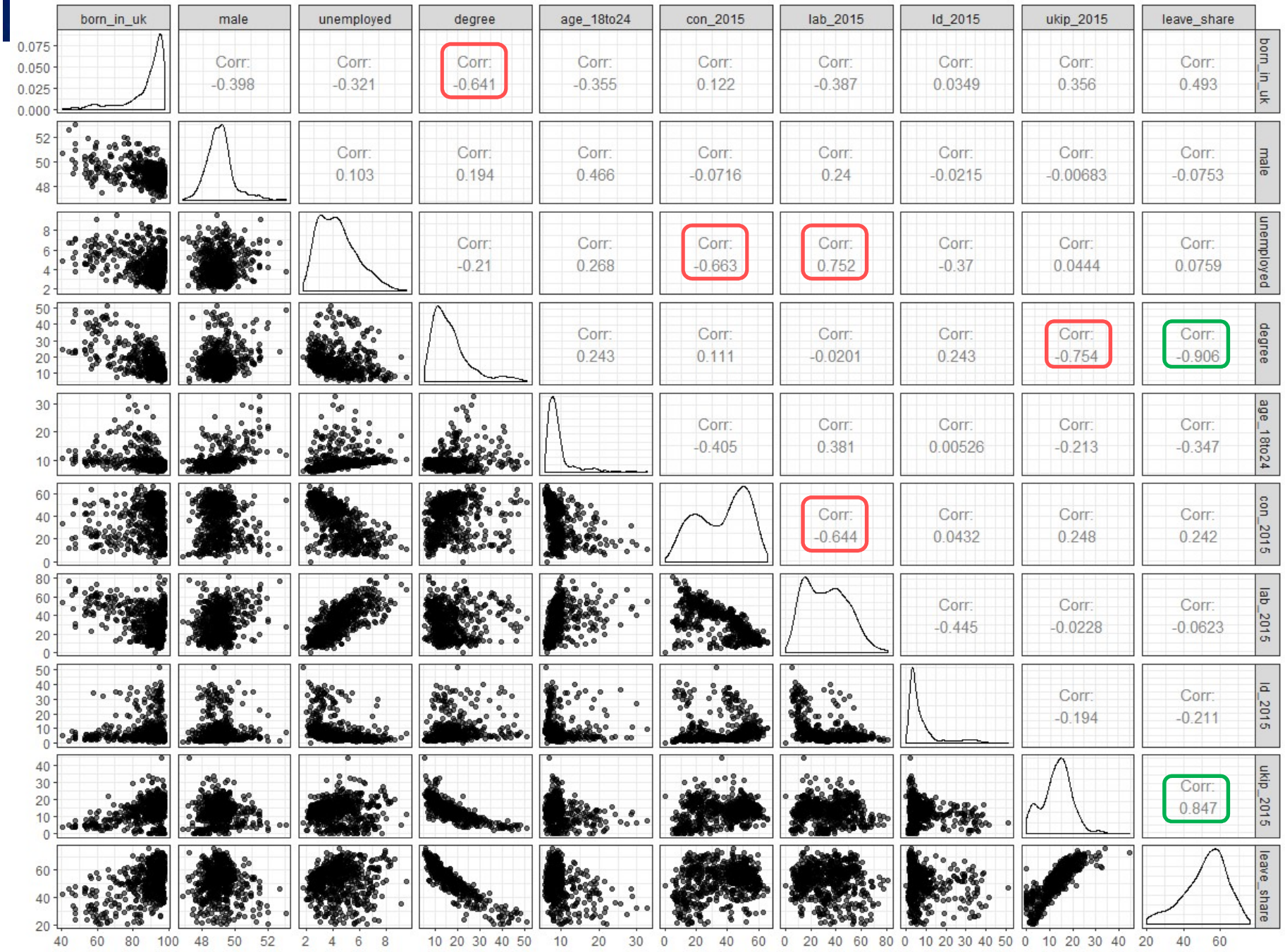


London Business School



London Business

Brexit correlation- scatterplot



London Business School

Brexit all predictors

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	6.091	11.174	0.545	0.586	-15.858	28.040
con_2015	0.237	0.023	10.440	0.000	0.192	0.281
lab_2015	0.129	0.025	5.132	0.000	0.080	0.179
ld_2015	0.133	0.028	4.761	0.000	0.078	0.188
ukip_2015	0.787	0.043	18.458	0.000	0.703	0.871
degree	-0.826	0.031	-26.909	0.000	-0.887	-0.766
age_18to24	-0.260	0.048	-5.459	0.000	-0.353	-0.166
born_in_uk	-0.007	0.021	-0.354	0.723	-0.048	0.033
male	0.743	0.204	3.646	0.000	0.343	1.144
unemployed	0.449	0.189	2.376	0.018	0.078	0.820

r_squared	adj_r_squared	mse	rmse	sigma	statistic	p_value	df
0.92	0.919	9.313516	3.051805	3.079	721.187	0	10

```
> vif(leave_everything)
  con_2015   lab_2015    ld_2015  ukip_2015     degree age_18to24 born_in_uk       male unemployed
  7.050112  10.847648   2.967263   3.402759   3.973118   1.764195   3.506365   1.514946   4.396018
```

Colinearity:

- Labour (lab_2015) has Variance Inflation Factor (VIF) > 10; Drop and rerun regression

```
> vif(leave_everything2)
  con_2015    ld_2015  ukip_2015     degree age_18to24 born_in_uk       male unemployed
  2.718999   1.318469   2.444450   3.938543   1.764052   3.489675   1.512837   3.898942
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	17.832	11.181	1.595	0.111	-4.130	39.793
con_2015	0.146	0.014	10.108	0.000	0.117	0.174
ld_2015	0.026	0.019	1.372	0.171	-0.011	0.063
ukip_2015	0.671	0.037	18.160	0.000	0.598	0.743
degree	-0.841	0.031	-26.910	0.000	-0.903	-0.780
age_18to24	-0.262	0.049	-5.386	0.000	-0.358	-0.166
born_in_uk	-0.015	0.021	-0.695	0.488	-0.056	0.027
male	0.704	0.208	3.382	0.001	0.295	1.114
unemployed	0.775	0.182	4.261	0.000	0.418	1.133

London Business School

Final brexit model and prediction

term <chr>	estimate <dbl>	std_error <dbl>	statistic <dbl>	p_value <dbl>	lower_ci <dbl>	upper_ci <dbl>
intercept	16.132	9.606	1.679	0.094	-2.736	35.001
con_2015	0.143	0.014	10.504	0.000	0.116	0.170
ukip_2015	0.661	0.036	18.188	0.000	0.589	0.732
degree	-0.827	0.025	-33.263	0.000	-0.876	-0.778
age_18to24	-0.261	0.049	-5.367	0.000	-0.356	-0.165
male	0.718	0.199	3.602	0.000	0.326	1.109
unemployed	0.767	0.135	5.678	0.000	0.502	1.033

r_squared <dbl>	adj_r_squared <dbl>	mse <dbl>	rmse <dbl>	sigma <dbl>	statistic <dbl>	p_value <dbl>	df <dbl>
0.916	0.915	9.790015	3.1289	3.148	1030.02	0	7

```
# Here are six imaginary constituencies, all with the same variables except
# education, which goes up by five in each row
imaginary_constituency3 <- tibble(con_2015 = 35,
                                  lab_2015 = 40,
                                  ld_2015 = 6,
                                  ukip_2015 = 15,
                                  degree = c(5, 10, 15, 20, 25, 30),
                                  age_18to24 = 30,
                                  born_in_uk = 80,
                                  male = 50,
                                  unemployed = 6)
# When we plug this multi-row data frame into predict(), it'll generate a
# prediction for each row
predict(final_model, newdata = imaginary_constituency3, interval = "prediction")

# We can also use broom::augment(). It's  essentially the same thing as predict(),
# but it adds the predictions and confidence intervals to the imaginary constituency
model_predictions <- broom::augment(final_model,
                                    newdata = imaginary_constituency3)
```

	con_2015	lab_2015	ld_2015	ukip_2015	degree	age_18to24	born_in_uk	male	unemployed
1	35	40	6	15	5	30	80	50	6
2	35	40	6	15	10	30	80	50	6
3	35	40	6	15	15	30	80	50	6
4	35	40	6	15	20	30	80	50	6
5	35	40	6	15	25	30	80	50	6
6	35	40	6	15	30	30	80	50	6

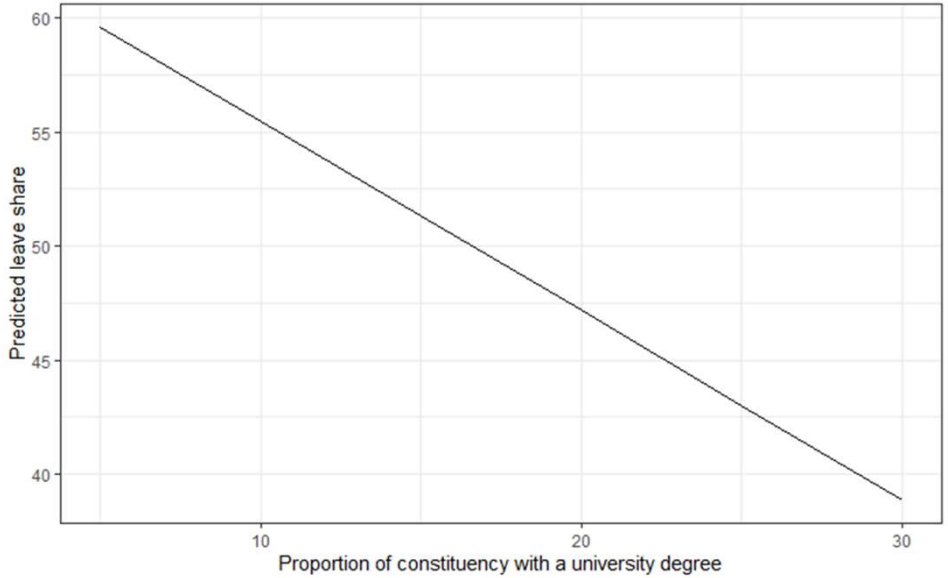
Final brexit model and prediction

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
intercept	16.132	9.606	1.679	0.094	-2.736	35.001
con_2015	0.143	0.014	10.504	0.000	0.116	0.170
ukip_2015	0.661	0.036	18.188	0.000	0.589	0.732
degree	-0.827	0.025	-33.263	0.000	-0.876	-0.778
age_18to24	-0.261	0.049	-5.367	0.000	-0.356	-0.165
male	0.718	0.199	3.602	0.000	0.326	1.109
unemployed	0.767	0.135	5.678	0.000	0.502	1.033

r_squared	adj_r_squared	mse	rmse	sigma	statistic	p_value	df
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0.916	0.915	9.790015	3.1289	3.148	1030.02	0	7

con_2015	lab_2015	ld_2015	ukip_2015	degree	age_18to24	born_in_uk	male	unemployed	.fitted
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
35	40	6	15	5	30	80	50	6	59.57640
35	40	6	15	10	30	80	50	6	55.44072
35	40	6	15	15	30	80	50	6	51.30504
35	40	6	15	20	30	80	50	6	47.16936
35	40	6	15	25	30	80	50	6	43.03368
35	40	6	15	30	30	80	50	6	38.89800

```
       fit      lwr      upr
1 59.57640 53.06736 66.08544
2 55.44072 48.95644 61.92500
3 51.30504 44.83638 57.77370
4 47.16936 40.70713 53.63159
5 43.03368 36.56865 49.49871
6 38.89800 32.42096 45.37504
```



London Business School

Model comparison

	(1)	(2)	(3)	(4)	(5)	(6)
(Intercept)	7.977 *	76.510 ***	36.685 ***	6.091	17.832	16.132
	(3.121)	(0.558)	(1.562)	(11.174)	(11.181)	(9.606)
born_in_uk	0.500 ***			-0.007	-0.015	
	(0.035)			(0.021)	(0.021)	
degree		-1.125 ***		-0.826 ***	-0.841 ***	-0.827 ***
		(0.022)		(0.031)	(0.031)	(0.025)
age_18to24		-0.456 ***		-0.260 ***	-0.262 ***	-0.261 ***
		(0.052)		(0.048)	(0.049)	(0.049)
con_2015			-0.018	0.237 ***	0.146 ***	0.143 ***
			(0.021)	(0.023)	(0.014)	(0.014)
lab_2015			-0.070 **	0.129 ***		
			(0.023)	(0.025)		
ld_2015			-0.128 ***	0.133 ***	0.026	
			(0.034)	(0.028)	(0.019)	
ukip_2015			1.472 ***	0.787 ***	0.671 ***	0.661 ***
			(0.039)	(0.043)	(0.037)	(0.036)
male				0.743 ***	0.704 ***	0.718 ***
				(0.204)	(0.208)	(0.199)
unemployed				0.449 *	0.775 ***	0.767 ***
				(0.189)	(0.182)	(0.135)
N	632	573	632	573	573	573
R2	0.243	0.843	0.726	0.920	0.916	0.916
logLik	-2348.258	-1646.561	-2027.534	-1452.367	-1465.468	-1466.662
AIC	4702.515	3301.122	4067.068	2926.734	2950.936	2949.325

*** p < 0.001; ** p < 0.01; * p < 0.05.

Do taller people have higher earnings?

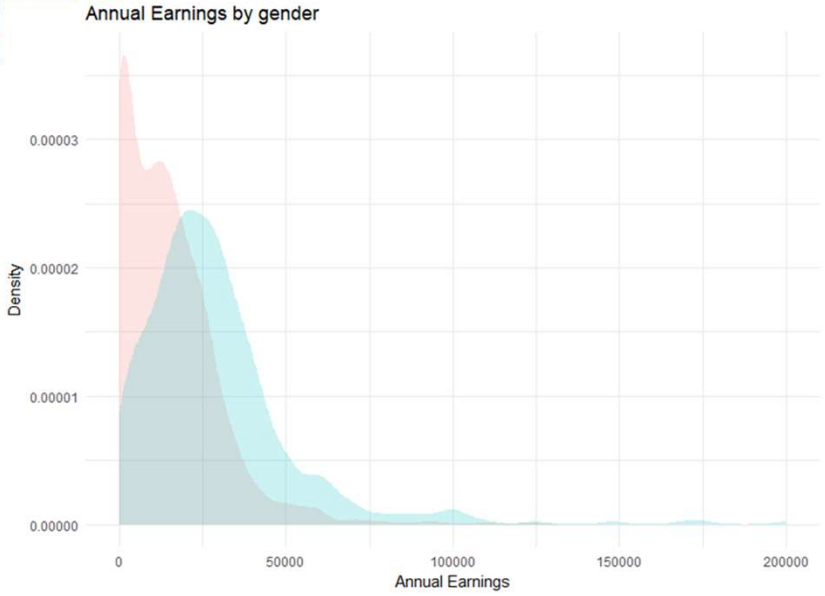
- Does height matter? If so, by how much?
- Data Source: *Work, Family, and Well-Being Survey* (Ross, 1990) http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/6666

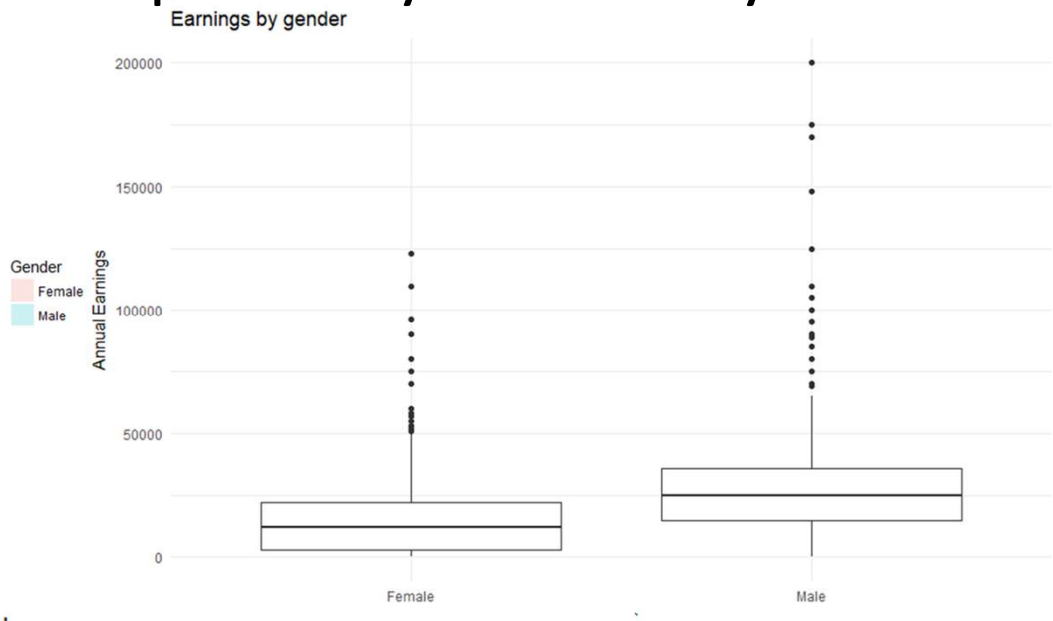
London Business School

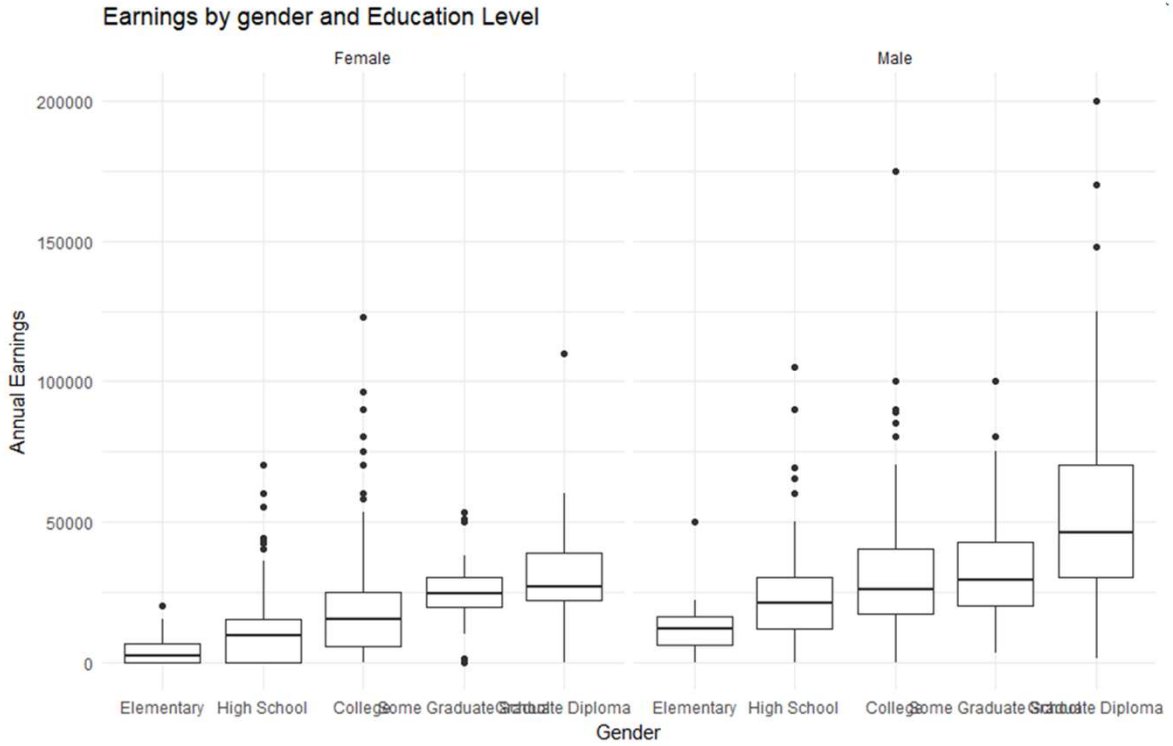
Inspect and Clean Data

```
> summary(height_earnings)
      earn            height          age                      ed_level                race          hispanic          gender
 Min.   :     0   Min.   :147.3   Min.   :18.00   College             :563   asian          :  17   Min.   :0.00000   Female:857
 1st Qu.:  6000   1st Qu.:162.6   1st Qu.:29.00   Elementary          : 46   black          : 126   1st Qu.:0.00000   Male  :519
 Median : 16020   Median :167.6   Median :38.00   Graduate Diploma    : 80   Native American:  11   Median :0.00000
 Mean   : 19985   Mean   :169.1   Mean   :41.22   High School         :617   other          :   5   Mean   :0.05233
 3rd Qu.: 28000   3rd Qu.:175.3   3rd Qu.:51.00   Some Graduate School: 70   white          :1217   3rd Qu.:0.00000
 Max.   :200000   Max.   :195.6   Max.   :89.00                                                      Max.   :1.00000
>
> skimr::skim(height_earnings)
── Data Summary ────────────────────────
                           Values
Name                       height_earnings
Number of rows             1376
Number of columns          7
_________________________
Column type frequency:
  factor                   3
  numeric                  4
_________________________
Group variables            None

── Variable type: factor ─────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 ed_level              0             1 FALSE          5 Hig: 617, Col: 563, Gra: 80, Som: 70
2 race                  0             1 FALSE          5 whi: 1217, bla: 126, asi: 17, Nat: 11
3 gender                0             1 FALSE          2 Fem: 857, Mal: 519

── Variable type: numeric ────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate      mean        sd    p0   p25    p50    p75   p100 hist
1 earn                  0             1 19985.     19768.       0  6000  16020  28000  200000 ▇▁▁▁▁
2 height                0             1   169.         9.66   147.  163.   168.   175.   196. ▁▃▇▅▂
3 age                   0             1    41.2       15.6     18    29     38     51     89 ▇▆▅▃▁
4 hispanic              0             1     0.0523     0.223    0     0      0      0      1 ▇▁▁▁▁
```

Exploratory Data Analysis- Plots





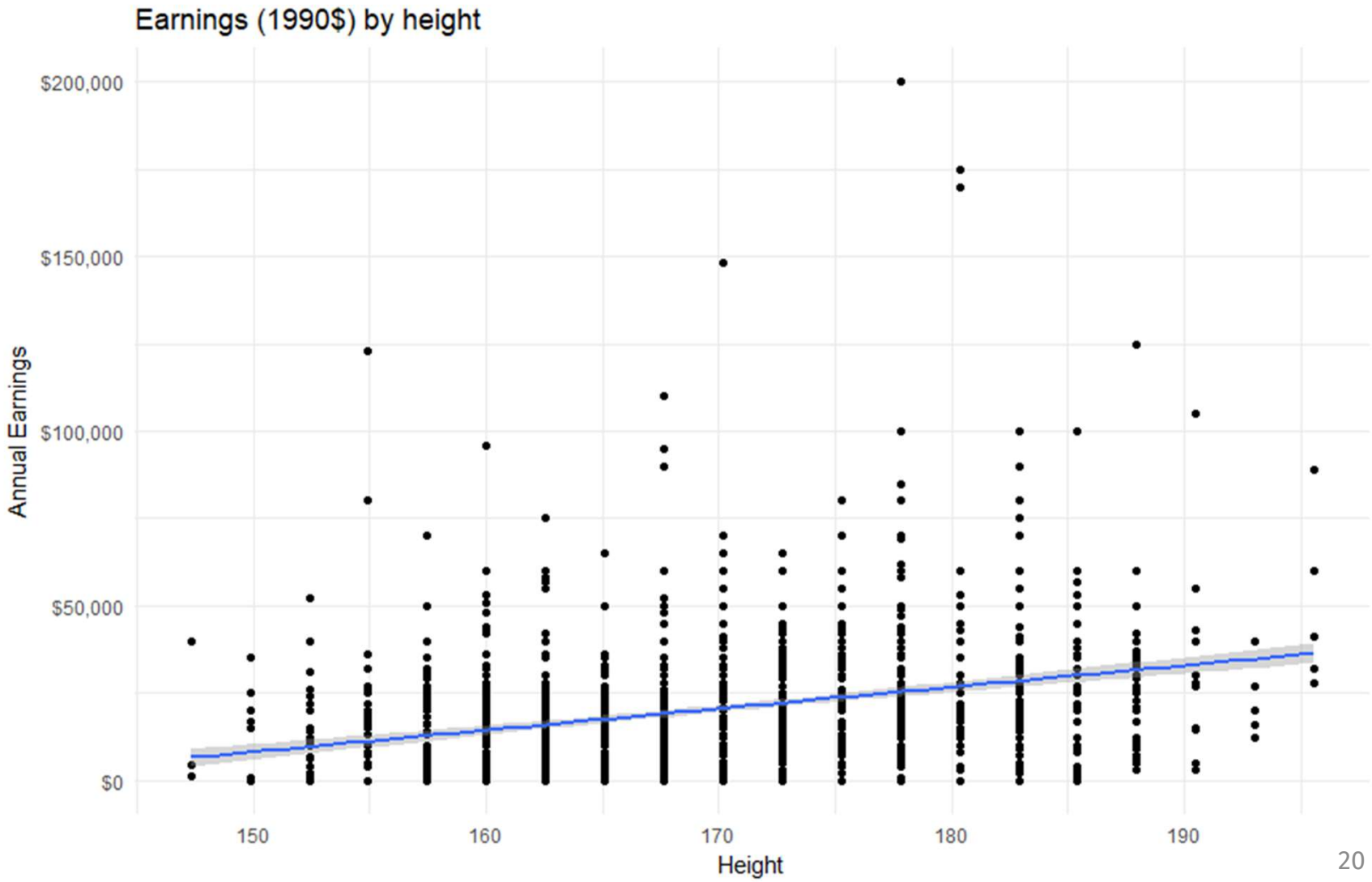


Earnings, height, and gender

- Data Source: *Work, Family, and Well-Being Survey* (Ross, 1990) http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/6666
- In the sample data, the average height was about 170cm and the average income about 20,000$
- I would like you to draw by hand a pair of axes representing earnings and height, and a point at (170,20000) which is the average height of adults in the US and their average earnings (in 1990).
- I now want you to draw a line through this central point with slope of roughly 620, carefully connecting the central point (170,20000) and two points (for heights of 150 and 190cm) on the regression line: (150, 20000 – 20 * 620) and (190, 20000 + 20 *620)).
- The regression line predicting earnings (Y) from height (X) is roughly
    - $(y - 20{,}000) = 620*(x - 170)$, or $y = -84{,}780 + 620*x$.
- Work in pairs and sketch a scatterplot of data that are consistent with this regression line.

Earnings on height

The survey may have issues of **response** and **measurement** error. Can you think of any?



London Business School

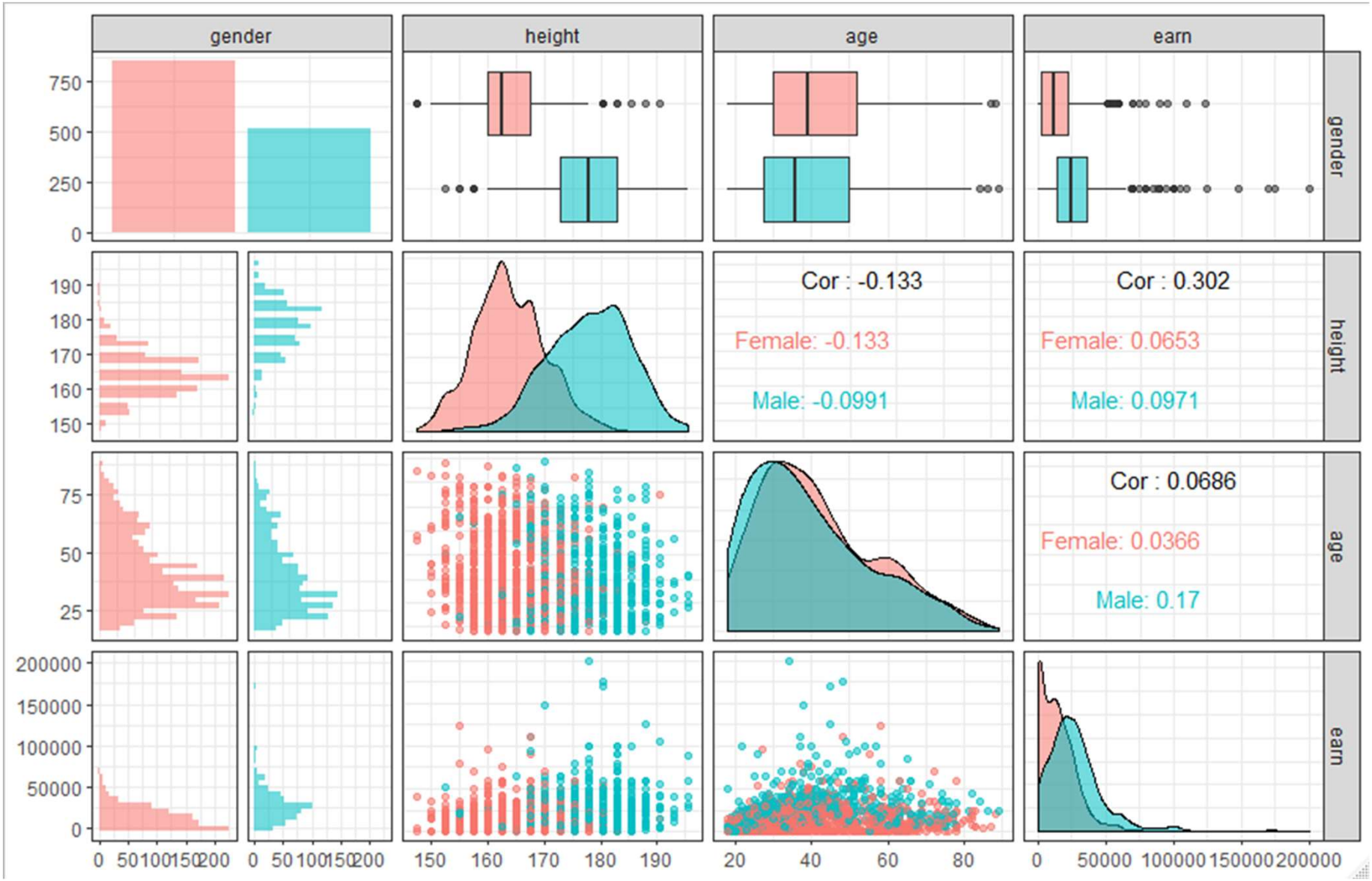
Earnings, height, and gender

- The regression slope ($ 619 per cm) is small but undeniably positive. Is the slope significant?
- How do we interpret the intercept in the regression equation? What is –84,634?
- Alternative way to write regression equation: $y = \bar{y} + b * (x - \bar{x}) = 20{,}000 + 619 * (x - 169)$
- How do you interpret the result that taller people have higher earnings?
- Are there any confounding variables? How do we control for them?

```
> model1 <- lm(earn~height, data=height_earnings)
> msummary(model1)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -84633.9     8911.3    -9.5   <2e-16
height         618.5       52.6    11.8   <2e-16

Residual standard error: 18800 on 1374 degrees of freedom
Multiple R-squared:  0.0914,    Adjusted R-squared:  0.0908
F-statistic:  138 on 1 and 1374 DF,  p-value: <2e-16
> confint(model1)
                2.5 %   97.5 %
(Intercept) -102115.1 -67152.8
height          515.3    721.7
```

London Business School

Correlation- Scatterplot Matrix



London Business School

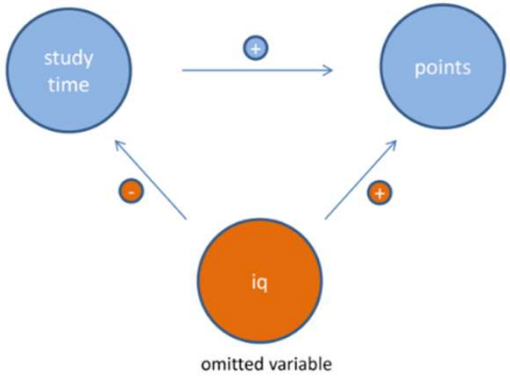
Salary on Height by Gender

Confounding variable bias (1/2)

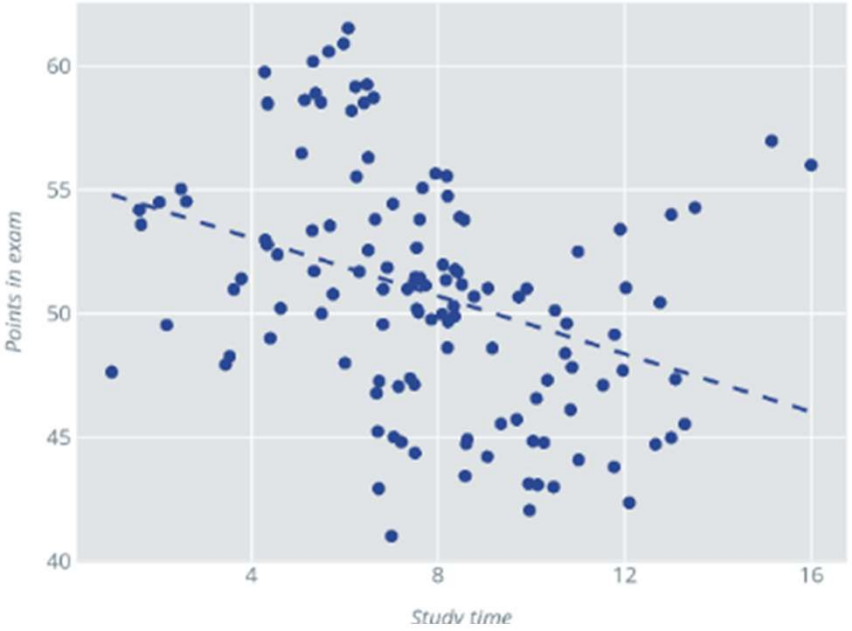
Confounding, or omitted, variable bias occurs when a variable not included in the model is correlated to both the dependent (Y) and the explanatory variable (X).

- Influences on the dependent variable (Y) which are not captured by the model are collected in the error term, which we want (or hope!) to be uncorrelated with the explanatory variable (X).
- This assumption is violated if we exclude determinants of the dependent variable (Y) which vary with the explanatory variable (X) .
- This might induce an estimation bias, i.e., the mean of the OLS estimator's sampling distribution no longer equals the true mean. This means we will wrongly estimate the effect on Y of a unit change in X, on average.

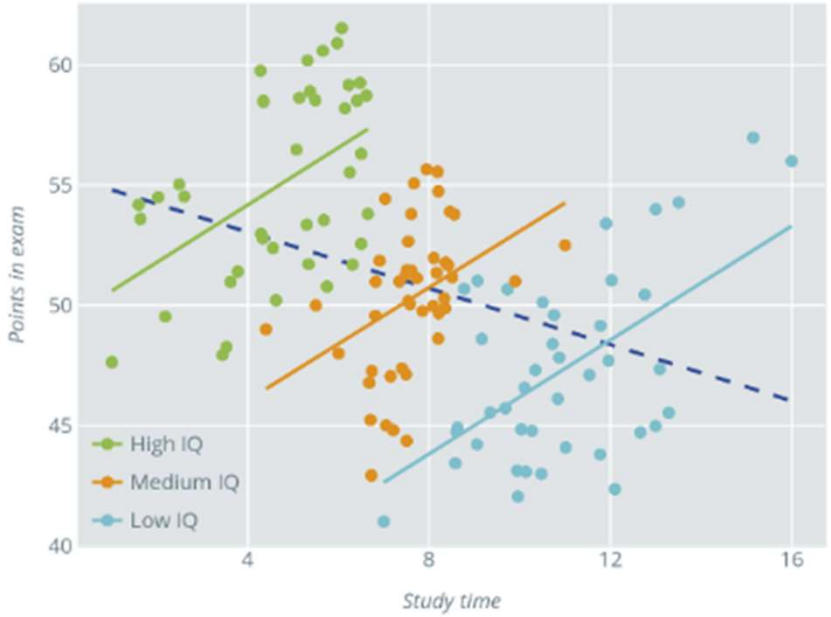
As an example, we would expect that the more effort someone puts in preparing for an exam, the higher the number of points scored in the exam.



Confounding variable bias (2/2)



If we just use *Study Time* as the only predictor, we find a negative relationship, which is counter-intuitive.



If we incorporate *IQ* as further predictor, we find the relationship between study time and points scored is positive and multiple regression allows us to estimate the positive effect of study time.

London Business School

Controlling for Gender

```
> model2 <- lm(earn~height+gender, data=height_earnings)
> msummary(model2)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -21334.0    11923.8   -1.79   0.0738
height         219.1       72.7    3.02   0.0026
genderMale   11280.8     1448.6    7.79  1.3e-14

Residual standard error: 18500 on 1373 degrees of freedom
Multiple R-squared:  0.13,      Adjusted R-squared:  0.129
F-statistic:  102 on 2 and 1373 DF,  p-value: <2e-16
> confint(model2)
                2.5 %  97.5 %
(Intercept) -44724.75  2056.7
height         76.56    361.7
genderMale   8439.19 14122.4
```

1. Comparing two people with the same height, what is the effect of being a man?
2. Is height predictive for differences in salary, even after controlling for sex?
3. Are the coefficients for height and gender statistically significant? Why?
4. What does the Residual Standard Error of 18500 indicate? What does the (adjusted) R^2 tell us? Why?
5. Draw plots of earnings vs. height, with separate regression lines for men and women. What earnings do you predict for a man of 169cm? For a woman of 169cm?

Extending the Earnings and Height model

- Besides *Height* and *Male* as explanatory variables, we also have data on *age*, education (5 categories), race (5 categories) and whether the respondents were Hispanic (1) or not (0).
- Before coding education and race as categorical variables, do you think there any issues with the sample? Why?

Sample Statistics on Education			
Education Level		N	%
1	Elementary	46	3.3%
2	High School	617	44.8%
3	College	563	40.9%
4	Some Graduate	70	5.1%
5	Graduate Diploma	80	5.8%

Sample Statistics on Race				
Race		N	%	US Census
1	white	1217	88.4%	72.4%
2	black	126	9.2%	12.6%
3	asian	17	1.2%	4.8%
4	native american	11	0.8%	0.9%
5	other	5	0.4%	
	Hispanics	72	5.2%	16.3%

Multiple regression model (1/2)

```
> model3 <- lm(earn ~ ., data=height_earnings)
> msummary(model3)
                    Estimate Std. Error t value     Pr(>|t|)
(Intercept)         -32958.8    12013.3   -2.74      0.00616
height                161.8       68.7    2.36      0.01857
age                   140.2       30.4    4.61 0.0000043022
ed_levelEd_2         8847.6     2669.1    3.31      0.00094
ed_levelEd_3        15823.0     2684.9    5.89 0.0000000048
ed_levelEd_4        20845.4     3283.6    6.35 0.0000000003
ed_levelEd_5        35572.3     3198.5   11.12       < 2e-16
raceblack             387.5     4434.8    0.09      0.93038
raceNative American  -3319.3     6611.7   -0.50      0.61573
raceother             2579.0     8748.2    0.29      0.76819
racewhite             2282.3     4201.1    0.54      0.58703
hispanic             -3569.5     2083.1   -1.71      0.08684
genderMale           11607.2     1349.0    8.60       < 2e-16

Residual standard error: 17000 on 1363 degrees of freedom
Multiple R-squared:  0.268,     Adjusted R-squared:  0.262
F-statistic: 41.6 on 12 and 1363 DF,  p-value: <2e-16
> confint(model3)
                      2.5 %  97.5 %
(Intercept)         -56525.38 -9392.3
height                 27.13   296.5
age                    80.58   199.7
ed_levelEd_2         3611.71 14083.6
ed_levelEd_3        10555.96 21090.1
ed_levelEd_4        14403.88 27286.9
ed_levelEd_5        29297.79 41846.8
raceblack           -8312.31  9087.4
raceNative American -16289.41  9650.9
raceother           -14582.41 19740.4
racewhite            -5958.95 10523.6
hispanic             -7655.93   516.9
genderMale            8960.91 14253.4
```

Not significant dummies for race. Drop them and re-model

When we have categorical variables, the one left out of the regression model is our 'baseline' with respect to that category. So for education, the first level was 'Elementary'. What is the meaning of the slope for *ed_levelEd_3* = 15823?

London Business School

Multiple regression model (2/2)

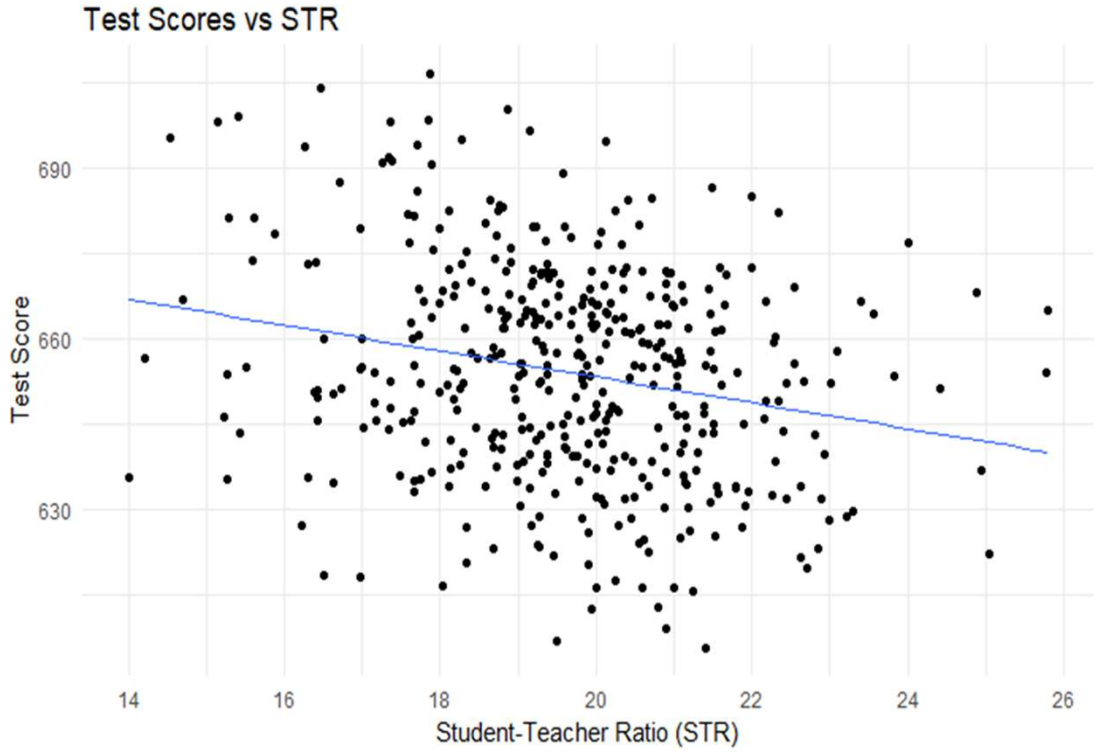
```
> model4 <- lm(earn ~ . -race -hispanic, data=height_earnings)
> msummary(model4)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -34238.6    11551.7   -2.96  0.00309
height          177.8       67.7    2.63  0.00868
age             144.8       30.2    4.80  1.8e-06
ed_levelEd_2   9204.7     2635.9    3.49  0.00049
ed_levelEd_3  16098.3     2656.5    6.06  1.8e-09
ed_levelEd_4  21201.5     3269.9    6.48  1.2e-10
ed_levelEd_5  35923.6     3165.1   11.35  < 2e-16
genderMale    11385.7     1335.9    8.52  < 2e-16

Residual standard error: 17000 on 1368 degrees of freedom
Multiple R-squared:  0.265,     Adjusted R-squared:  0.262
F-statistic: 70.6 on 7 and 1368 DF,  p-value: <2e-16
> confint(model4)
                2.5 %   97.5 %
(Intercept)  -56899.6 -11577.6
height           45.1    310.6
age              85.6    204.0
ed_levelEd_2   4033.9  14375.6
ed_levelEd_3  10887.1  21309.6
ed_levelEd_4  14787.0  27616.1
ed_levelEd_5  29714.7  42132.6
genderMale     8765.0  14006.4
```

1. Is height predictive for differences in salary, even after controlling for sex, age, and education?
2. What is the effect of education on earnings?

Do smaller classes lead to higher test scores?

A prominent policy proposal to improve learning in elementary schools is to reduce class size

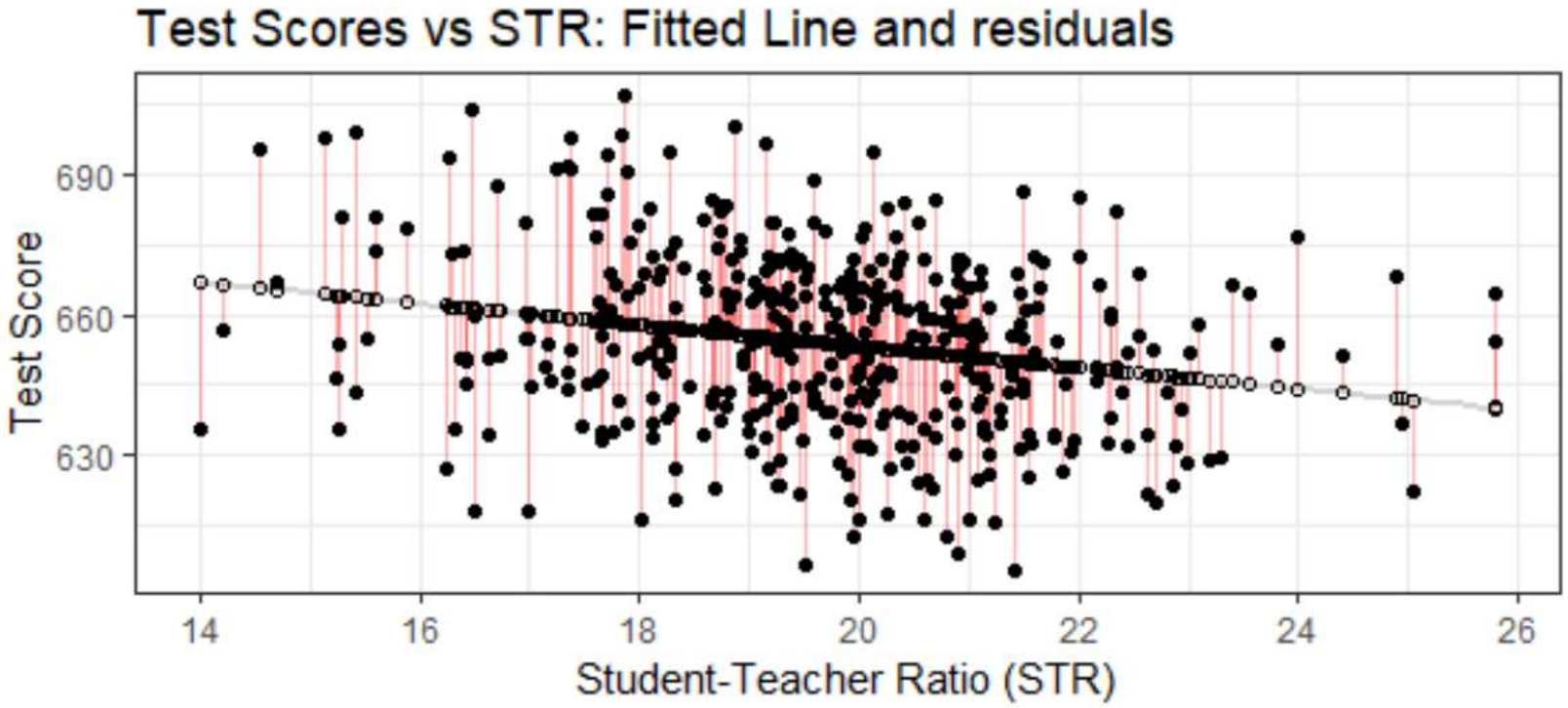
- Is the beneficial effect on basic learning (*three Rs*: reading, writing and arithmetic) of smaller classes large or small?
- Is it possible that smaller class size has no effect on basic learning?
- How can we isolate the effect of changes in class size form changes in other factors, like the socio-economic background of the students?
- How can we predict likely test scores for a district?
- Data on 420 California school districts in 1999



London Business School

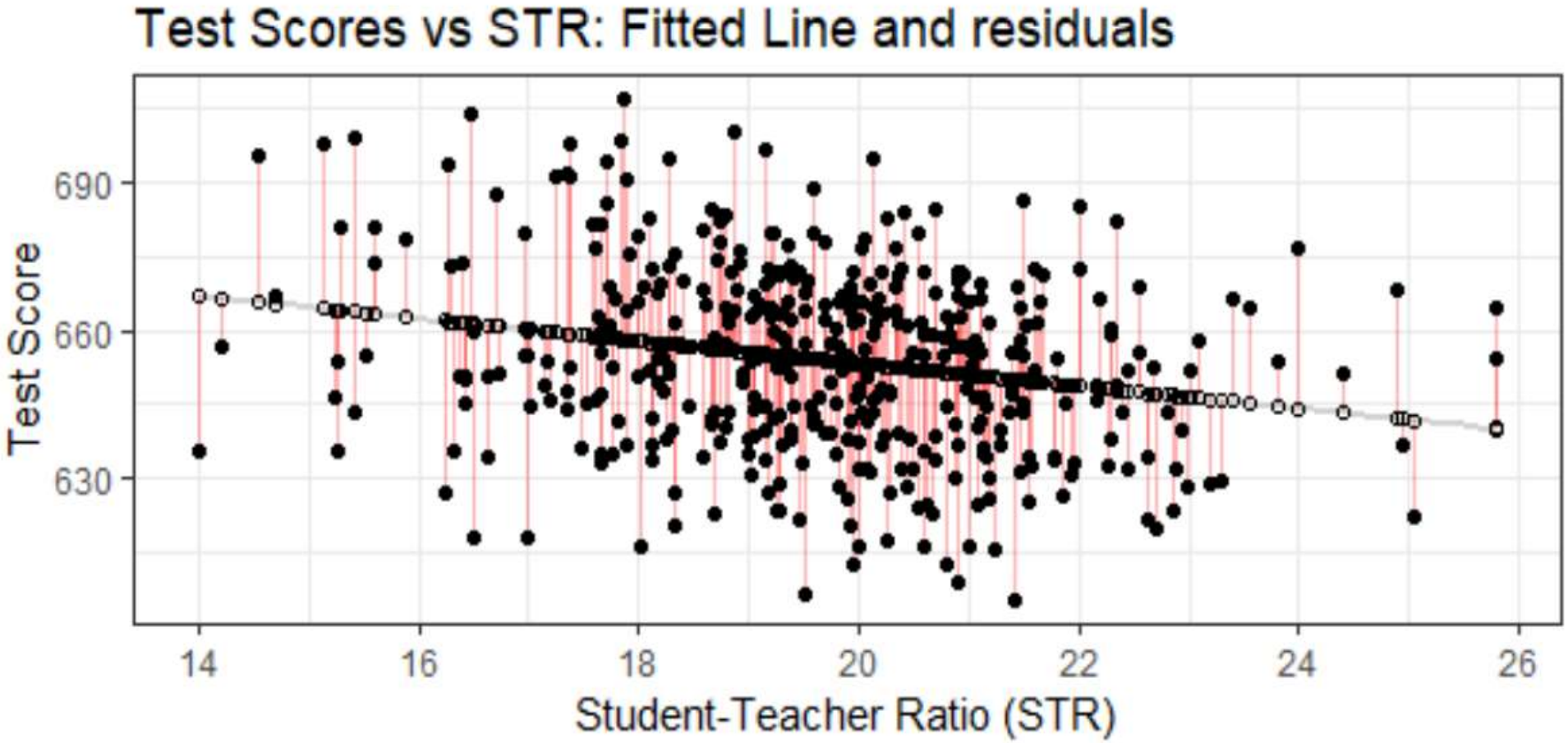
Plotting fitted line and residuals

```
# Plot fitted line and residuals
ggplot(model1_aug, aes(x=str, y=test_score)) +
  geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +  # Plot regression slope
  geom_segment(aes(xend = str, yend = .fitted), color="red", alpha = 0.3) +  # alpha to fade lines
  geom_point() +
  geom_point(aes(y = .fitted), shape = 1) +
  labs(
      x="Student-Teacher Ratio (STR)",
      y="Test Score",
      title = "Test Scores vs STR: Fitted Line and residuals") +
  theme_bw()  # Add theme for cleaner look +
  NULL
```



London Business School

Plotting fitted line and residuals



Interpretation of estimated slope and intercept

test_score = 698.93 – 2.28* str

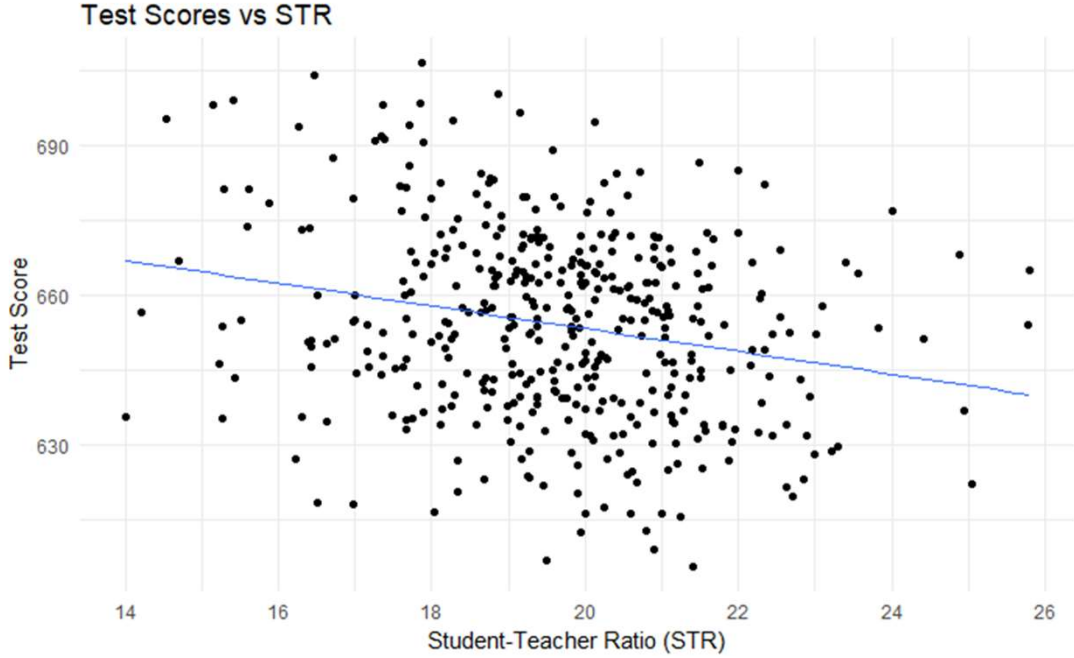
Districts with one more student per teacher have test scores that are 2.28 points lower (significantly different from zero as *t-Stat*= -4.75).

The CI for STR's slope is between [-3.22,-1.34]

The intercept of 698.93 tells us that districts with zero students per teacher would have predicted score of about 699

- This interpretation is non-sensical! It extrapolates the line outside the data range, so the intercept here is not meaningful

```
> model1 <- lm(test_score~str,data=CA_test_scores)
> msummary(model1)
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)   698.93       9.47   73.82 < 0.0000000000000002
str            -2.28       0.48   -4.75            0.0000028

Residual standard error: 18.6 on 418 degrees of freedom
Multiple R-squared:  0.0512,    Adjusted R-squared:  0.049
F-statistic: 22.6 on 1 and 418 DF,  p-value: 0.00000278
> anova(model1)
Analysis of Variance Table

Response: test_score
           Df Sum Sq Mean Sq F value    Pr(>F)
str         1   7794    7794    22.6 0.0000028
Residuals 418 144315     345
> confint(model1)
               2.5 %   97.5 %
(Intercept) 680.323 717.5428
str          -3.223  -1.3366
```



London Business School

Prediction and Confidence Intervals

Suppose we wanted to find out what ***TestScores*** would be for

1. the next, random school district with an STR of 22
2. Schools districts with an average STR of 22

Prediction intervals tell you where you can expect to see the **next data point** sampled. The key point is that the prediction interval tells you about the distribution of values, not the uncertainty in determining the population mean.

Confidence intervals are intervals for the model and tell you about how well you have determined **the mean Y**. The key point is that the confidence interval tells you about the likely location of the true population parameters of the model.

Prediction intervals must account for both the uncertainty in knowing the value of the population mean, plus the error/noise along the line. So a prediction interval is always wider than a confidence interval.

Prediction and Confidence Intervals

To calculate a **point prediction**, or best guess, substitute the given values of the *X*s into the estimated regression equation.

- To measure the accuracy of the point predictions, calculate a **standard error** for each prediction.
  - Standard error of prediction for an individual *Y*, *for a given value of x:*

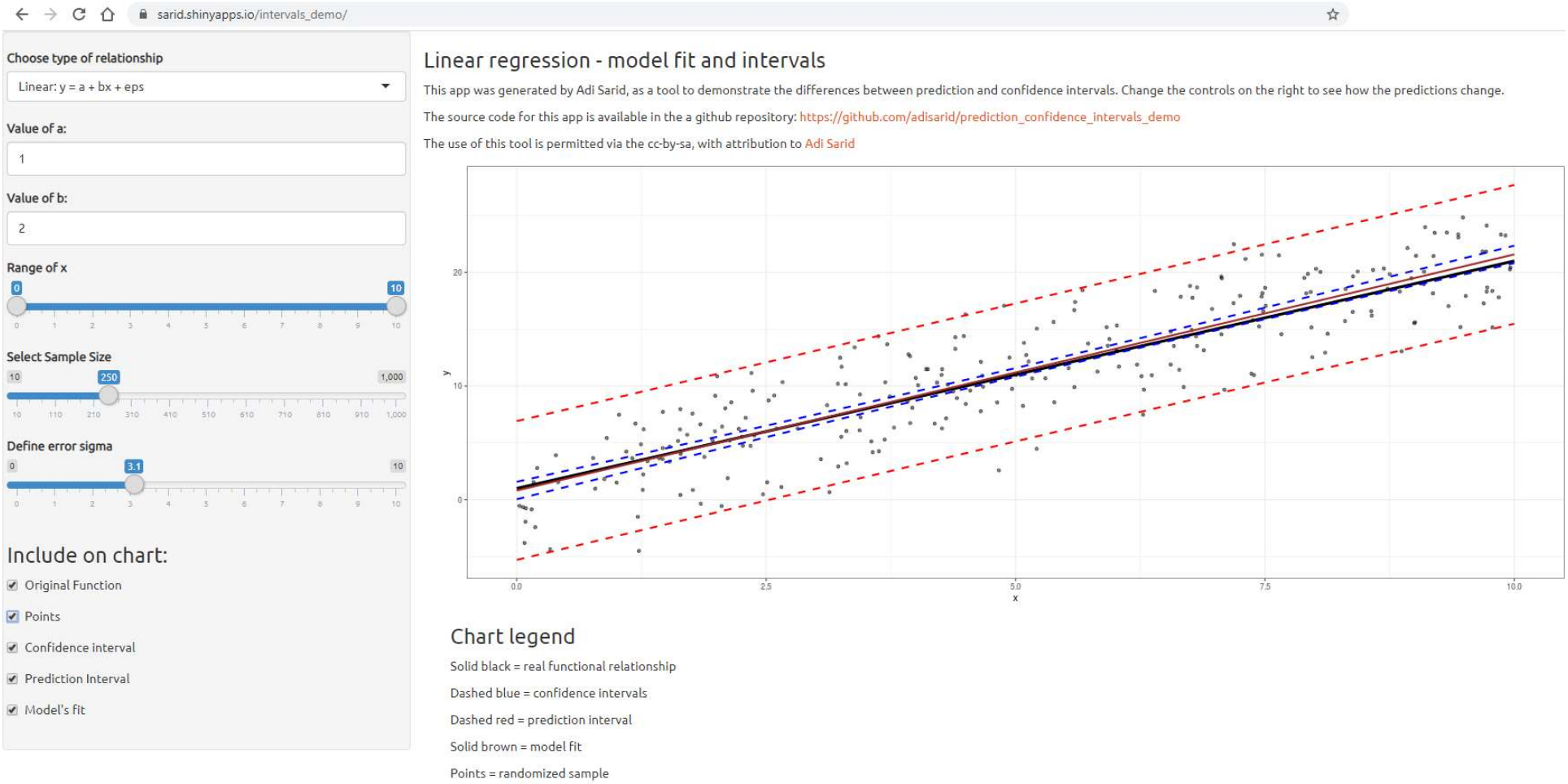
$$s_{\text{ind}} = s_e \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \simeq s_e$$

    - This error is approximately equal to the regression standard error.
  - Standard error of prediction for the mean *Y*, *for a given X or E(Y|X), Confidence interval:*

$$s_{\text{mean}} = s_e \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \simeq s_e/\sqrt{n}$$

    - This error is approximately equal to the regression standard error divided by the square root of the sample size.

Prediction and Confidence Intervals

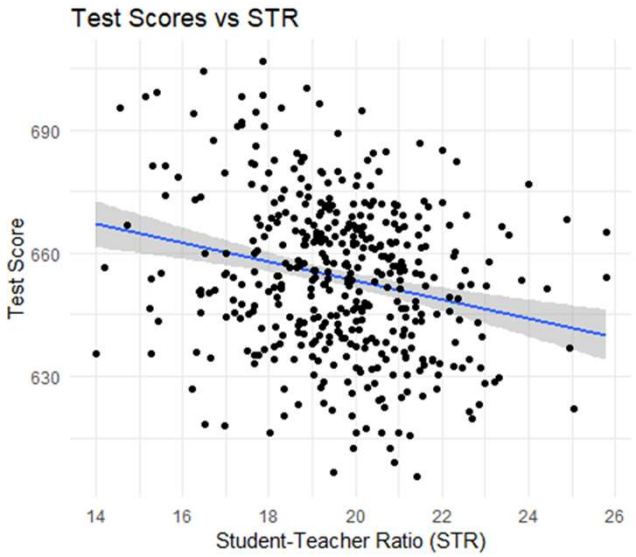


Source: https://sarid.shinyapps.io/intervals_demo/

Confidence interval for E[Y|X]

- The confidence interval for E[Y|X] is narrowest at the mean value of X and the band widens as the distance from the mean of X increases.

```
# scatter plot with linear model
ggplot(CA_test_scores, aes(x=STR, y=testScore)) +
  geom_smooth(method='lm', se=TRUE, level=0.95) +
  geom_point()+
  labs(x="Student-Teacher Ratio (STR)", y="Test Score", title = "Test Scores vs STR") +
  theme_minimal() +
  NULL
```



These standard errors can be used to calculate a 95% prediction interval for an individual value and a 95% confidence interval for a mean value.

- Go out a *t*-multiple (typically 1.96) of the relevant SE on either side of the point prediction.
- The term *prediction interval* (rather than confidence interval) is used for an individual value because an individual value of *Y* is not a population *parameter*.

London Business School

Using *broom:augment*

Instead of viewing the coefficients, you might be interested in the fitted values, residuals and SEs for each of the original points in the regression. For this, use broom::augment()

```
library(broom)
model1_aug <- augment(model1)
```

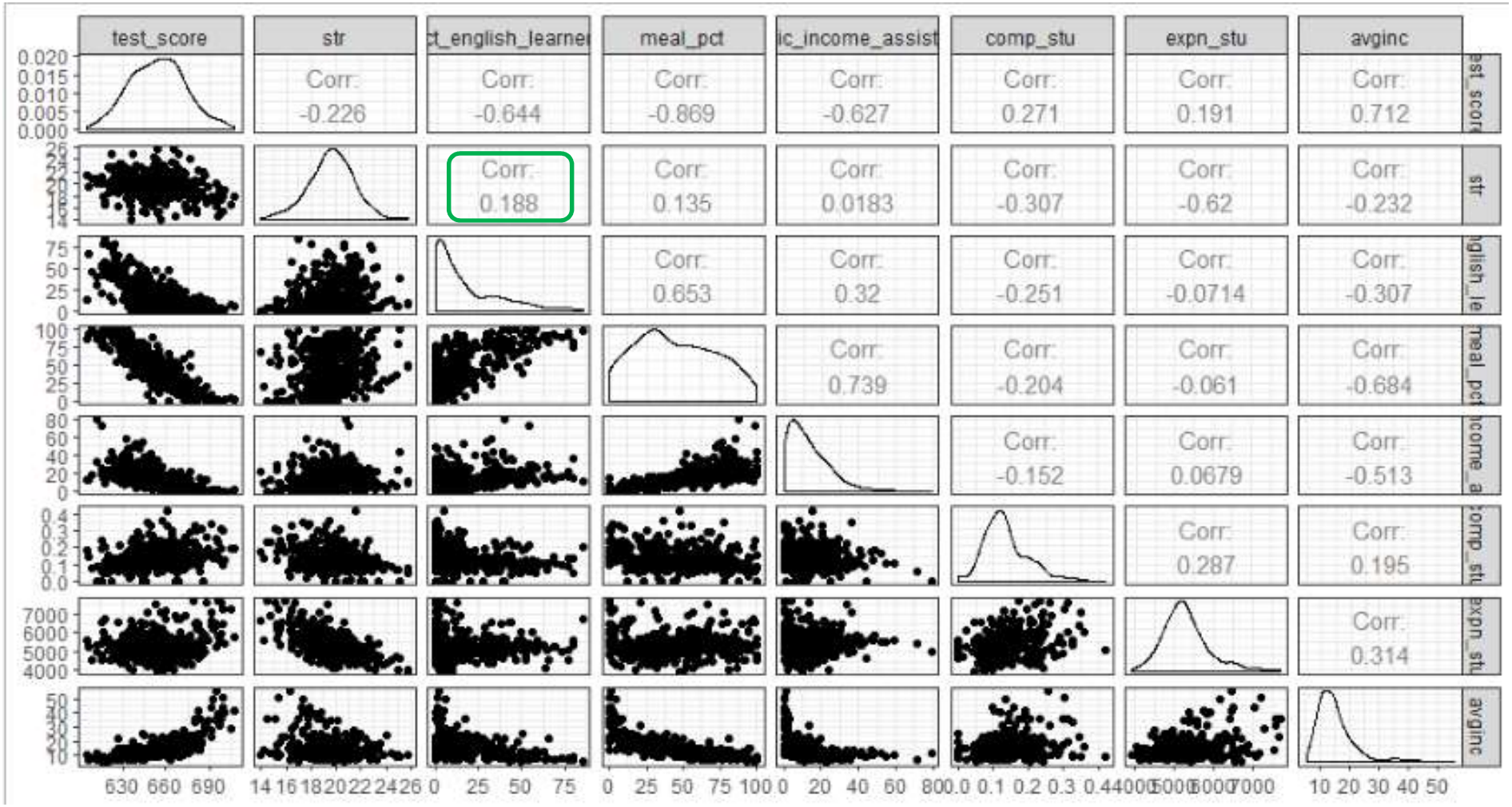
	test_score	str	.fitted	.se.fit	.resid	.hat	.sigma	.cooksd	.std.resid
1	690.80	17.890	658.15	1.23593	32.652608	0.0044244	18.534	0.00689246706324	1.76121524
2	661.20	21.525	649.86	1.28040	11.339155	0.0047485	18.595	0.00089266208745	0.61171055
3	643.60	18.697	656.31	1.01334	-12.706863	0.0029742	18.593	0.00069962884576	-0.68488368
4	647.70	17.357	659.36	1.42208	-11.661995	0.0058575	18.594	0.00116733312734	-0.62947749
5	640.85	18.671	656.37	1.01895	-15.515903	0.0030072	18.588	0.00105479859269	-0.83630120
6	605.55	21.406	650.13	1.24094	-44.580805	0.0044603	18.474	0.01295314515021	-2.40464125
7	606.75	19.500	654.48	0.90916	-47.726691	0.0023941	18.455	0.00793561664480	-2.57165968
8	609.00	20.894	651.30	1.08807	-42.298369	0.0034291	18.487	0.00894625256864	-2.28034834
9	612.50	19.947	653.46	0.91854	-40.956776	0.0024438	18.495	0.00596584701198	-2.20693099
10	612.65	20.806	651.50	1.06517	-38.850274	0.0032862	18.505	0.00723065788580	-2.09430790
11	615.75	21.238	650.51	1.18731	-34.764166	0.0040831	18.525	0.00720516205143	-1.87478706
12	616.30	21.000	651.06	1.11696	-34.756977	0.0036136	18.525	0.00636796902129	-1.87395770
13	616.30	20.600	651.97	1.01687	-35.668901	0.0029950	18.521	0.00555152127675	-1.92252827
14	616.30	20.008	653.32	0.92367	-37.018053	0.0024712	18.514	0.00492844719967	-1.99472267

SE for predicting mean Y for a given X, or E(Y|X)

SE for predicting an individual *Y*

Adding further regressors

Before adding ***pct_english_learners*** as a regressor, we need to check whether our explanatory variables (***str*** and ***pct_english_learners***) are highly correlated



London Business School

Adding further regressors

```
> model2 <- lm(test_score ~ str + pct_english_learners,data=CA_test_scores)
> msummary(model2)
                    Estimate Std. Error t value             Pr(>|t|)
(Intercept)         686.0323     7.4113    92.6 <0.0000000000000002
str                  -1.1013     0.3803    -2.9                0.004
pct_english_learners -0.6498     0.0393   -16.5 <0.0000000000000002

Residual standard error: 14.5 on 417 degrees of freedom
Multiple R-squared:  0.426,     Adjusted R-squared:  0.424
F-statistic:  155 on 2 and 417 DF,  p-value: <0.0000000000000002
> confint(model2)
                       2.5 %    97.5 %
(Intercept)         671.46407 700.60044
str                  -1.84880  -0.35379
pct_english_learners -0.72711  -0.57244
```

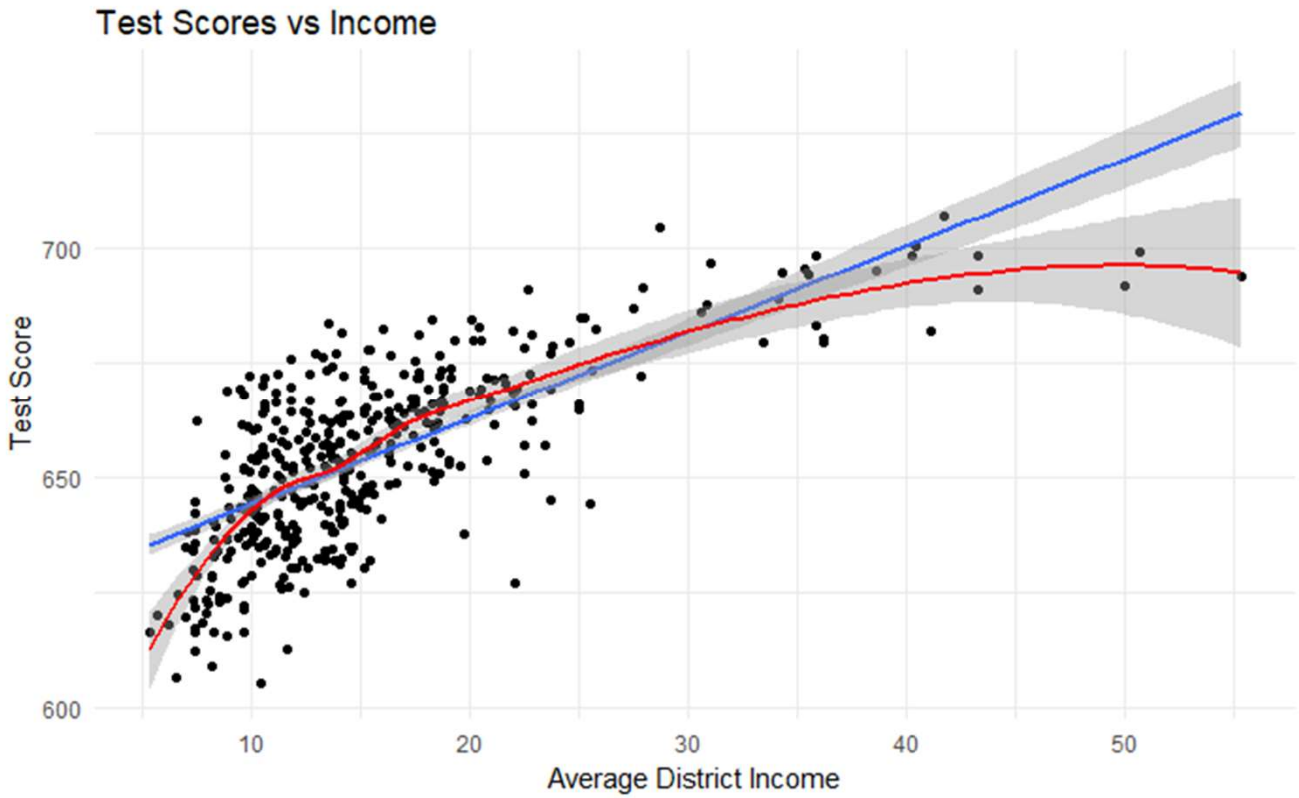
TestScore = 686 – 1.1* str -0.65*pct_English_learners

- Adjusted R^2 improves to 0.424 (from 0.05)
- Residual SE 14.5 (compared to 18.6)
- No Collinearity, as correlation(str, *pct_english_learners* =0.188)

Non-linear relationships

Relationship between income and test score is not linear. Two approaches we can take

1. Polynomials in X
   - The population regression function is approximated by a quadratic, cubic, or higher-degree polynomial
2. Logarithmic transformations
   - Y and/or X is transformed by taking its logarithm
   - this gives a “percentages” interpretation that makes sense in many applications



Non-linear relationships

Run a regression with a polynomial, *averageIncome* and *averageIncome*^{2}

```
> model_quadratic <- lm(test_score~avginc + I(avginc^2),data=CA_test_scores)
> msummary(model_quadratic)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 607.30175    3.04622  199.36  < 2e-16 ***
avginc        3.85099    0.30426   12.66  < 2e-16 ***
I(avginc^2)  -0.04231    0.00626   -6.76  4.7e-11 ***

Residual standard error: 12.7 on 417 degrees of freedom
Multiple R-squared:  0.556,     Adjusted R-squared:  0.554
F-statistic:  261 on 2 and 417 DF,  p-value: <2e-16
```

Test Score = 607.3+ 3.85**averageIncome*- 0.042**averageIncome*^{2}

Compute "effects" for different values of X (income)

- Predicted change in TestScore for a change in income from 5K to 6K

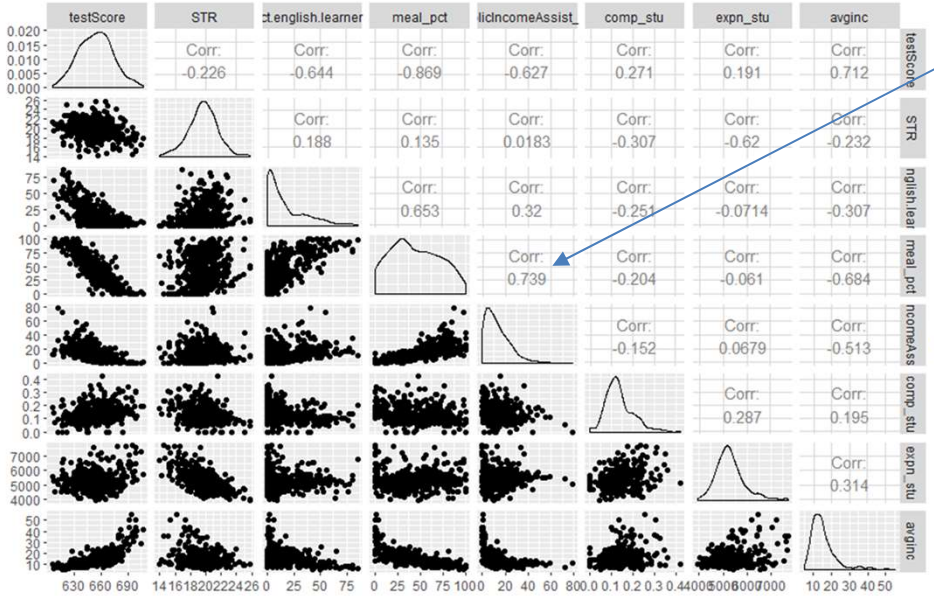
  $TS = 607.3+3.85\times6-0.042\times6^2 - (607.3+3.85\times5-0.042\times5^2) = 3.4$
- Predicted change in Test Score for a change in income from 25K to 26K → TestScore = 1.7
- Predicted change in Test Score for a change in income from 45K to 46K → TestScore = 0.0

- The effect of a change in income is greater at low than high income

Comparing models

Comparison of models

	(1)	(2)	(3)	(4)	(5)
(Intercept)	698.933	686.032	700.150	697.999	700.392
	(9.467)	(7.411)	(4.686)	(6.024)	(4.698)
str	-2.280	-1.101	-0.998	-1.308	-1.014
	(0.480)	(0.380)	(0.239)	(0.307)	(0.240)
pct_english_learners		-0.650	-0.122	-0.488	-0.130
		(0.039)	(0.032)	(0.033)	(0.034)
meal_pct			-0.547		-0.529
			(0.022)		(0.032)
public_income_assist_pct				-0.790	-0.048
				(0.053)	(0.061)
#observations	420	420	420	420	420
R squared	0.051	0.426	0.775	0.629	0.775
Adj. R Squared	0.049	0.424	0.773	0.626	0.773
Residual SE	18.581	14.464	9.080	11.654	9.084



- Model 5 contains two predictors (meal_pct, public_income_assist_pct) that are highly correlated (0.739)
- Collinearity, i.e., predictors that are highly linearly related, can cause problems in estimating the regression coefficients
- Need to check whether the VIF *(Variance Inflation Factor)* is <5. using the ***car::vif()***, we get

```
> vif(model5)
     str pct_english_learners        meal_pct public_income_assist_pct
  1.0444                1.9623          3.8705                  2.4765
```

- While no VIF is > 5, model 3 gives us similar results (R^2, SE), is parsimonious, and avoids collinearity

London Business School

Final model and prediction (1/2)

```
> mosaic::msummary(final_model)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         700.1500     4.6857  149.42  < 2e-16 ***
str                  -0.9983     0.2388   -4.18  3.5e-05 ***
pct_english_learners -0.1216     0.0323   -3.76  0.00019 ***
meal_pct             -0.5473     0.0216  -25.34  < 2e-16 ***

Residual standard error: 9.08 on 416 degrees of freedom
Multiple R-squared:  0.775,     Adjusted R-squared:  0.773
F-statistic:  476 on 3 and 416 DF,  p-value: <2e-16
> # Here are six imaginary districts, all with the same variables except
> # meal_pct, which goes up by 10% in each row
> imaginary_district <- tibble(str = 22,
+                              pct_english_learners = 20,
+                              meal_pct =  c(5, 15, 25, 35, 45, 55))
> imaginary_district
# A tibble: 6 x 3
    str pct_english_learners meal_pct
  <dbl>                <dbl>    <dbl>
1    22                   20        5
2    22                   20       15
3    22                   20       25
4    22                   20       35
5    22                   20       45
6    22                   20       55
```

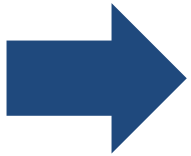
London Business School

Final model and prediction (2/2)

```
> # When we plug this multi-row data frame into predict(), it'll generate a
> # prediction for each row
> predict(final_model, newdata = imaginary_district, interval = "prediction")
     fit    lwr    upr
1 673.02 655.02 691.02
2 667.55 649.58 685.51
3 662.07 644.14 680.00
4 656.60 638.69 674.51
5 651.13 633.22 669.03
6 645.65 627.75 663.56
> # We can also get a confidence interval for Expected value of Y, given these X's
> predict(final_model, newdata = imaginary_district, interval = "confidence")
     fit    lwr    upr
1 673.02 670.68 675.36
2 667.55 665.53 669.56
3 662.07 660.34 663.80
4 656.60 655.08 658.12
5 651.13 649.72 652.53
6 645.65 644.24 647.07
> # We can also use broom::augment(). It's  essentially the same thing as predict(),
> # but it adds the predictions and SEs to the imaginary school district
> model_predictions <- broom::augment(final_model,
+                                     newdata = imaginary_district)
> model_predictions <- model_predictions %>%
+   mutate(
+     lower = .fitted - 1.96 * .se.fit,
+     upper = .fitted + 1.96 * .se.fit
+   )
> model_predictions
# A tibble: 6 x 7
    str pct_english_learners meal_pct .fitted .se.fit lower upper
  <dbl>                <dbl>    <dbl>   <dbl>   <dbl> <dbl> <dbl>
1    22                   20        5    673.   1.19   671.  675.
2    22                   20       15    668.   1.02   666.  670.
3    22                   20       25    662.   0.881  660.  664.
4    22                   20       35    657.   0.773  655.  658.
5    22                   20       45    651.   0.715  650.  653.
6    22                   20       55    646.   0.720  644.  647.
```

London Business School

Contents



- Further regression examples
- Interaction variables
- Regression Diagnostics

London Business School

Pay discrimination

- The Federal Upminster Bank is facing a pay-discrimination suit. The charge is that its female employees receive substantially smaller salaries than its male employees. A sample of the bank's employee database is listed in the file *bank_salaries.csv*
- Data set includes the following variables for each of the 208 employees of the bank: Education (5 categories), Job Grade (6 categories), years_bank (years with this bank), years_prior (years of previous work experience), age, gender, pc_job (categorical yes/no), Salary.

	gender	years_bank	years_prior	age	pc_job	education	job_grade	salary
1	Male	3	1	26	No	Ed_3	Grade 1	32000
2	Female	14	1	38	No	Ed_1	Grade 1	39100
3	Female	12	0	35	No	Ed_1	Grade 1	33200
4	Female	8	7	40	No	Ed_2	Grade 1	30600
5	Male	3	0	28	No	Ed_3	Grade 1	29000
6	Female	3	0	24	No	Ed_3	Grade 1	30500
7	Female	4	0	27	No	Ed_3	Grade 1	30000
8	Male	8	2	33	No	Ed_3	Grade 1	27000
9	Female	4	0	62	No	Ed_1	Grade 1	34000
10	Female	9	0	31	No	Ed_3	Grade 1	29500
11	Female	9	2	34	No	Ed_3	Grade 1	26800
12	Female	8	8	37	No	Ed_2	Grade 1	31300
13	Female	9	0	37	No	Ed_2	Grade 1	31200
14	Female	10	6	58	No	Ed_2	Grade 1	34700
15	Female	4	0	33	No	Ed_3	Grade 1	30000

Do the data provide evidence that females are discriminated against in terms of salary?

Bank Salaries – Model 0, the mean

- In the absence of any explanatory variables, our best guess for someone's salary would be the mean
- The sample mean = 39922
- To give a CI for the mean, we would calculate the standard error

$$SE = \frac{11256}{\sqrt{208}} = 780$$

```
> favstats(~salary, data=bank_salaries)
   min    Q1 median    Q3   max  mean    sd   n missing
 26700 33000  37000 44000 97000 39922 11256 208       0
>
> model0 <- lm(salary ~ 1, data=bank_salaries)
> msummary(model0)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    39922        780    51.1   <2e-16

Residual standard error: 11300 on 207 degrees of freedom
```

- Rather than summary statistics, we can run ***lm(salary ~ 1)*** and we get the same results.

Bank Salaries – Model 1

- Let us calculate summary statistics of salary by gender.
- The difference in mean salaries is 45505 – 37210 = 8296

```
> favstats(salary ~ gender, data=bank_salaries)
  gender   min    Q1 median    Q3   max  mean    sd   n missing
1 Female 26800 32575  35450 41550 61800 37210  6711 140       0
2   Male 26700 34375  42500 48625 97000 45505 15843  68       0
```

- Besides summary statistics, we can run ***lm(salary ~ gender)*** and we get the same results.
- The intercept of 37210 is the mean female salary and the slope of 8296 is the difference in mean salaries between men and women

```
> model1 <- lm(salary ~ gender, data=bank_salaries)
> msummary(model1)
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)    37210        894    41.6    < 2e-16
genderMale      8296       1564     5.3 0.00000029
```

London Business School

Bank Salaries – Model 2

```
> model2 <- lm(salary ~ ., data=bank_salaries)
> msummary(model2)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      27135.46    2455.28   11.05  < 2e-16
genderMale        2554.47    1011.97    2.52    0.012
years_bank         515.58      97.98    5.26  3.8e-07
years_prior        167.73     140.44    1.19    0.234
age                 -8.96      57.70   -0.16    0.877
pc_jobYes         4922.85    1473.82    3.34    0.001
educationEd_2     -485.55    1398.66   -0.35    0.729
educationEd_3      527.91    1357.52    0.39    0.698
educationEd_4      285.18    2404.73    0.12    0.906
educationEd_5     2690.80    1620.89    1.66    0.099
job_gradeGrade 2  1564.50    1185.77    1.32    0.189
job_gradeGrade 3  5219.36    1262.39    4.13  5.3e-05
job_gradeGrade 4  8594.83    1496.02    5.75  3.5e-08
job_gradeGrade 5 13659.41    1874.27    7.29  7.9e-12
job_gradeGrade 6 23832.39    2799.89    8.51  4.7e-15

Residual standard error: 5650 on 193 degrees of freedom
Multiple R-squared:  0.765,     Adjusted R-squared:  0.748
F-statistic: 44.9 on 14 and 193 DF,  p-value: <2e-16
```

Throw all the variables in the model. Are they all significant?

Interaction Variables

- Interaction variables are needed when there is reason to believe that the effect of one independent variable X1 depends on the value of another independent variable X2.
- General Equation with No Interaction: $Y_i = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 D + error$
- When you include only a dummy variable *(D)* in a regression equation you are allowing the intercepts of the two lines to differ, but you are forcing the lines to be parallel (*Parallel Slopes* Model).
- To be more realistic, you might want to allow them to have different slopes. You can do this by including an interaction variable.
  - An interaction variable is the *product* of two explanatory variables, *X1* and *X2*. We form an interaction variable that is the product of *X1*X2*
  - We have estimated the effect of *X1* and *X2* independently. With the interaction variable *X1*X2* we want to see whether there is an 'additional effect'

Interactions between a continuous and a categorical variable
Interactions between two binary variables
Interactions between two continuous variables

Interaction Variables (2/2)

Interactions Between a Continuous and a Categorical (Dummy) Variable

- The baseline model is $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 D + error$
- A multiple regression model that allows us to estimate the average difference of gender pay holding working experience (X1) constant as well as the average effect on earnings of a change in working experience holding gender (D) constant (*ceteris paribus*).
- By adding the interaction term X1*D we allow the effect of an additional year of work experience to differ between individuals with different genders.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 D + \beta_3 (X_1 * D) + error$$

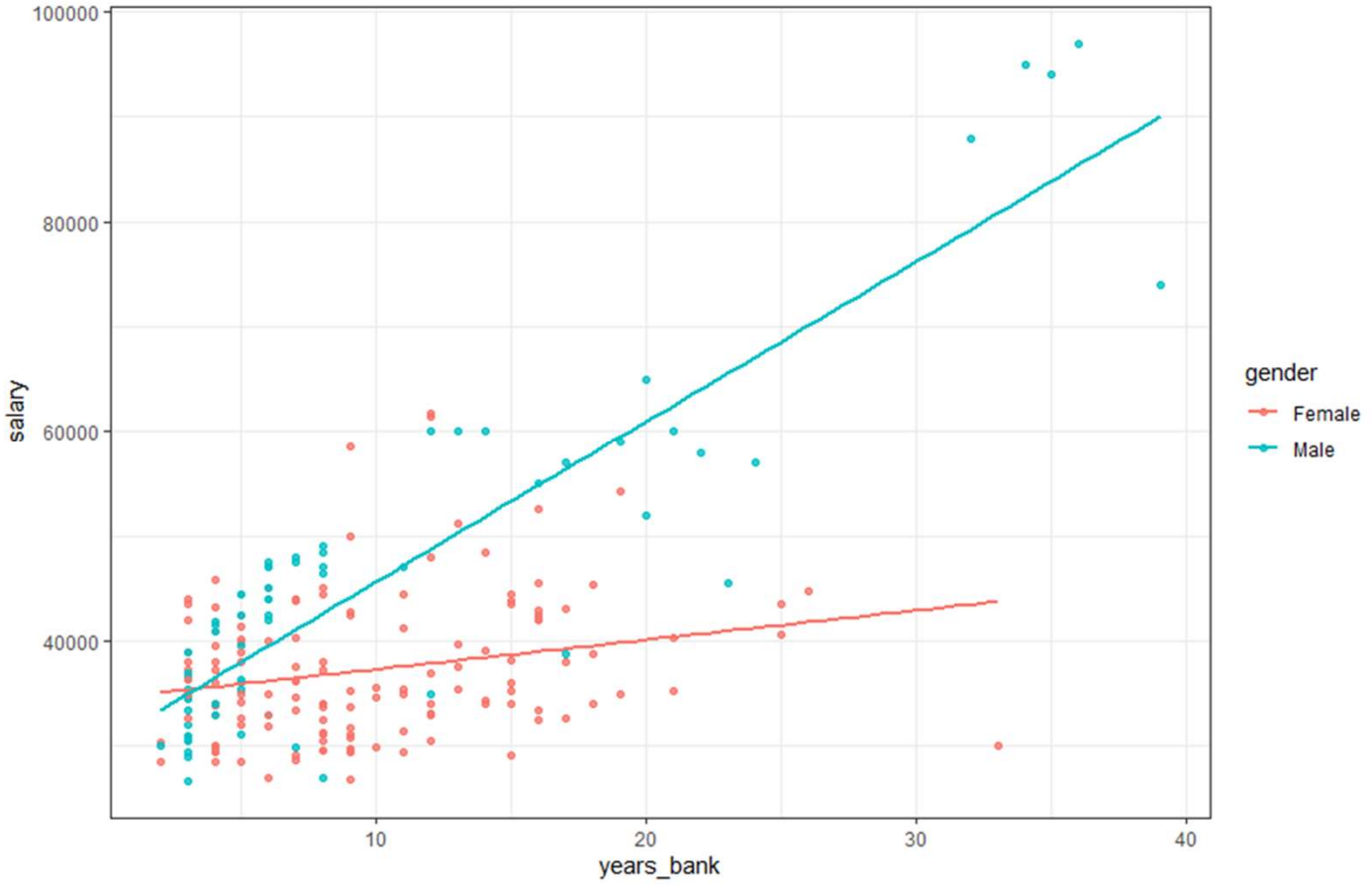
- Here, β3 is the expected difference in the effect of an additional year of work experience for men vs women.
- To specify a model where X, D, and the interaction X*D are all used in R, use

```
lm(y ~ X*D, data = mydata)
```

Interaction Variables: *gender*years_bank*

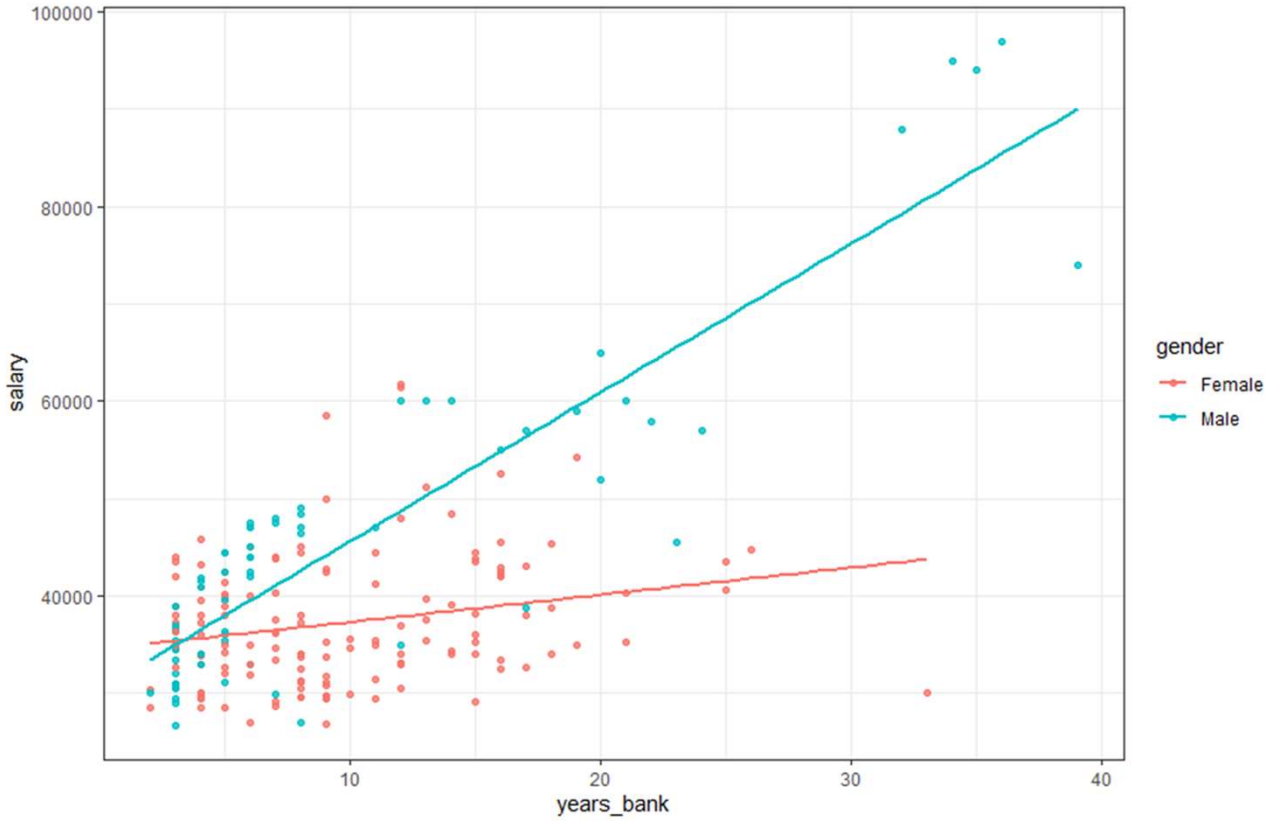
Interactions Between a Continuous (*years_bank*) and a Categorical (*gender*) Variable

- If we have reasons to believe that the annual salary increase depends not just on time passed, but also on your gender, then we introduce the interaction variable
- If this were a parallel slopes model, we would estimate the effect of time (*years_bank*) and then shift the line up/down to adjust for gender.
- However, when we plot salary on years_bank, and colour the points by gender, we see the two lines have fairly different slopes



Bank Salaries – Model 3

```
> model3 <- lm(salary ~ years_bank*gender , data=bank_salaries)
> msummary(model3)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)              34528       1138   30.34   <2e-16
years_bank                 280        102    2.73   0.0068
genderMale                -4098       1666   -2.46   0.0147
years_bank:genderMale      1248        137    9.13   <2e-16

Residual standard error: 6820 on 204 degrees of freedom
Multiple R-squared:  0.639,     Adjusted R-squared:  0.633
F-statistic:  120 on 3 and 204 DF,  p-value: <2e-16
```



- The average starting salary is about 34.5K
- Each year of experience increases salary by 280.
- Being male decreases your salary by 4098 …
- …BUT for each year of experience in the bank men's salaries go up by a further 1248
- So $ effect of an extra year is
  - 280 for women
  - 280+1248=1528 for men

```
> model4 <- lm(salary ~ years_bank*gender + education, data=bank_salaries)
> msummary(model4)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           29679.7     1444.2   20.55  < 2e-16
years_bank              426.5       91.8    4.65  6.1e-06
genderMale            -4898.7     1454.1   -3.37  0.00091
educationEd_2           546.5     1418.1    0.39  0.70035
educationEd_3          3587.3     1287.4    2.79  0.00584
educationEd_4          5862.9     2346.6    2.50  0.01328
educationEd_5          9428.1     1337.3    7.05  2.8e-11
years_bank:genderMale  1029.9      121.9    8.45  6.1e-15

Residual standard error: 5940 on 200 degrees of freedom
Multiple R-squared:  0.731,     Adjusted R-squared:  0.722
F-statistic: 77.8 on 7 and 200 DF,  p-value: <2e-16
>
>
> model5 <- lm(salary ~ years_bank*gender + education + job_grade, data=bank_salaries)
> msummary(model5)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           31690.9     1323.8   23.94  < 2e-16
years_bank              107.0       85.7    1.25    0.214
genderMale            -6066.1     1267.5   -4.79  3.4e-06
educationEd_2          -675.1     1204.7   -0.56    0.576
educationEd_3           447.3     1147.8    0.39    0.697
educationEd_4           525.1     2109.3    0.25    0.804
educationEd_5          1946.1     1394.6    1.40    0.164
job_gradeGrade 2       2245.4     1034.4    2.17    0.031
job_gradeGrade 3       5552.1     1098.5    5.05  9.9e-07
job_gradeGrade 4       9970.3     1314.6    7.58  1.3e-12
job_gradeGrade 5      13235.2     1631.4    8.11  5.4e-14
job_gradeGrade 6      14928.1     2695.7    5.54  9.8e-08
years_bank:genderMale  1002.9      119.1    8.42  7.9e-15

Residual standard error: 4990 on 195 degrees of freedom
Multiple R-squared:  0.815,     Adjusted R-squared:  0.804
F-statistic: 71.6 on 12 and 195 DF,  p-value: <2e-16
>
>
> model6 <- lm(salary ~ years_bank*gender + job_grade, data=bank_salaries)
> msummary(model6)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           32167.6      987.5   32.57  < 2e-16
years_bank               49.8       78.4    0.64    0.526
genderMale            -6063.3     1266.3   -4.79  3.3e-06
job_gradeGrade 2       2596.5     1010.1    2.57    0.011
job_gradeGrade 3       6221.4      998.2    6.23  2.7e-09
job_gradeGrade 4      11072.0     1172.6    9.44  < 2e-16
job_gradeGrade 5      14946.6     1340.2   11.15  < 2e-16
job_gradeGrade 6      17097.4     2390.7    7.15  1.6e-11
years_bank:genderMale  1021.1      118.7    8.60  2.4e-15

Residual standard error: 4990 on 199 degrees of freedom
Multiple R-squared:  0.811,     Adjusted R-squared:  0.803
F-statistic:  107 on 8 and 199 DF,  p-value: <2e-16
```

- Which is your preferred model and why?
- Are women discriminated against?

London Business School

Go to **www.menti.com** and use the code **41 70 04 5**

Mentimeter

In 2019, I went at least once to

London Business School

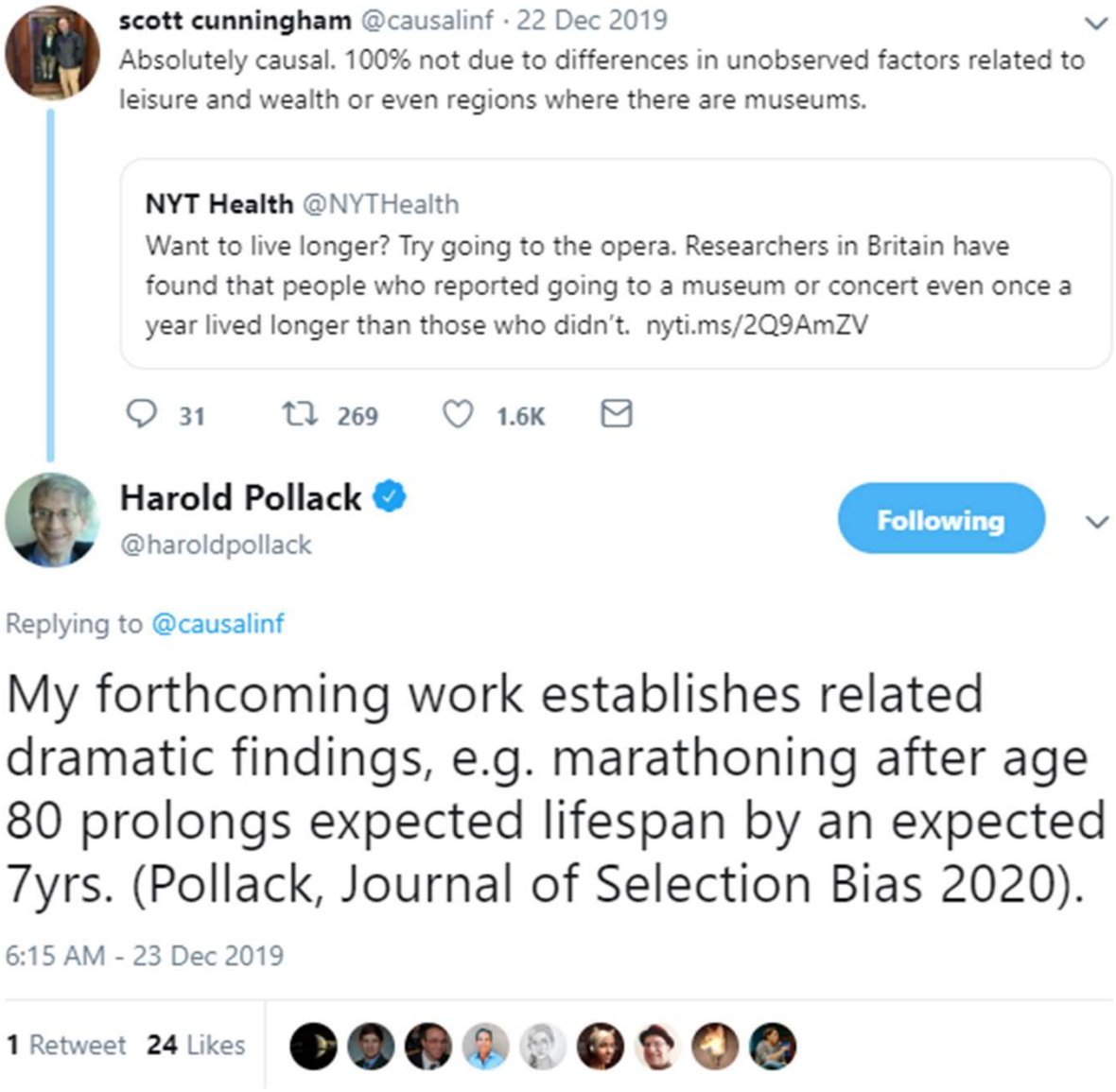
Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

nytimes.com

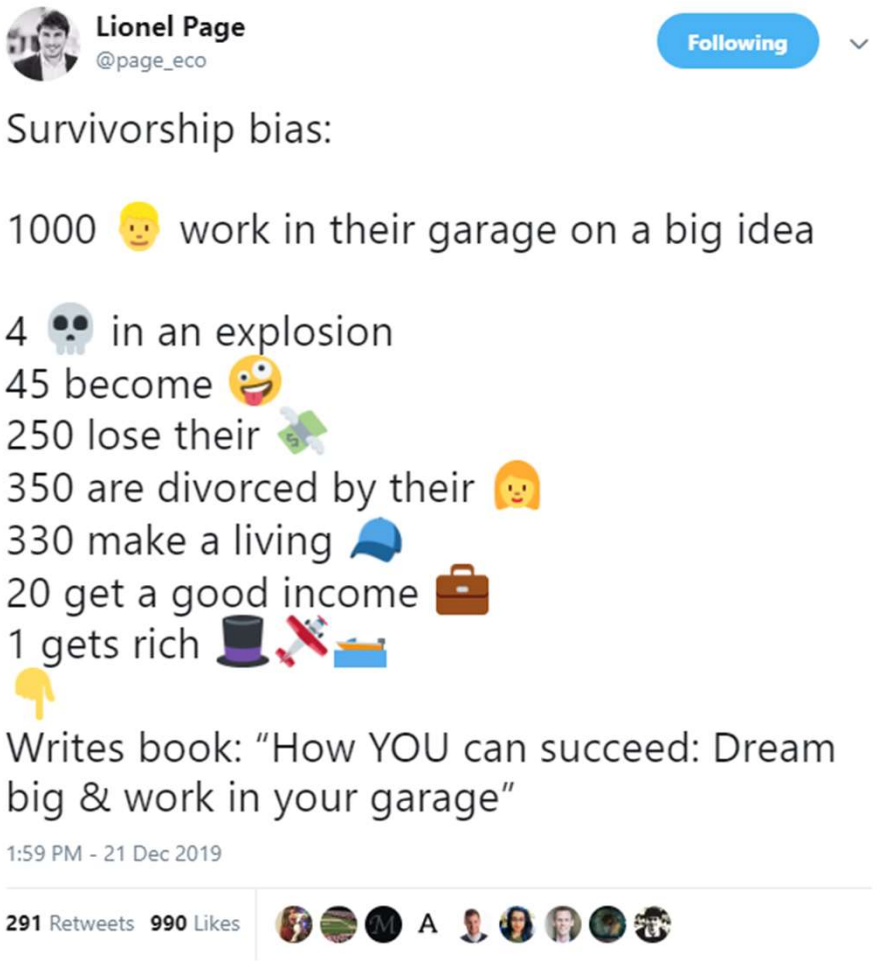
Source: https://www.nytimes.com/2019/12/22/us/arts-health-effects-ucl-study.html

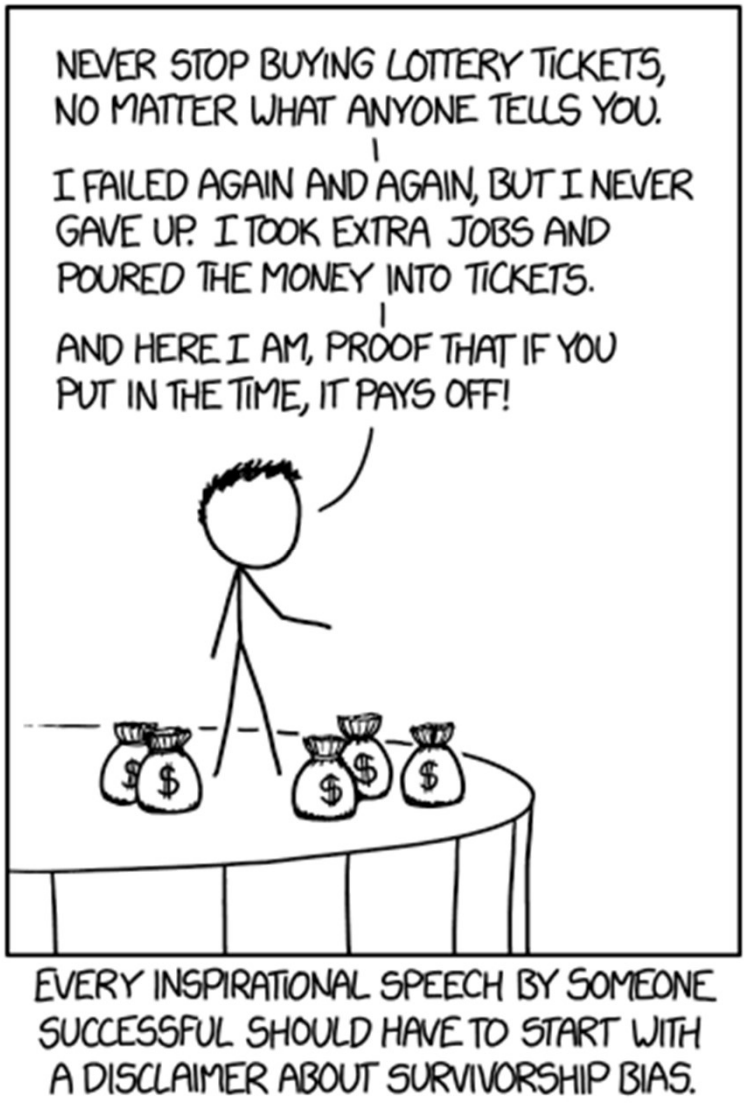
Selection bias



Source: https://twitter.com/haroldpollack/status/1208964401027526656

Survivorship bias





Source: https://twitter.com/page_eco/status/1208356430757056516

https://xkcd.com/1827/

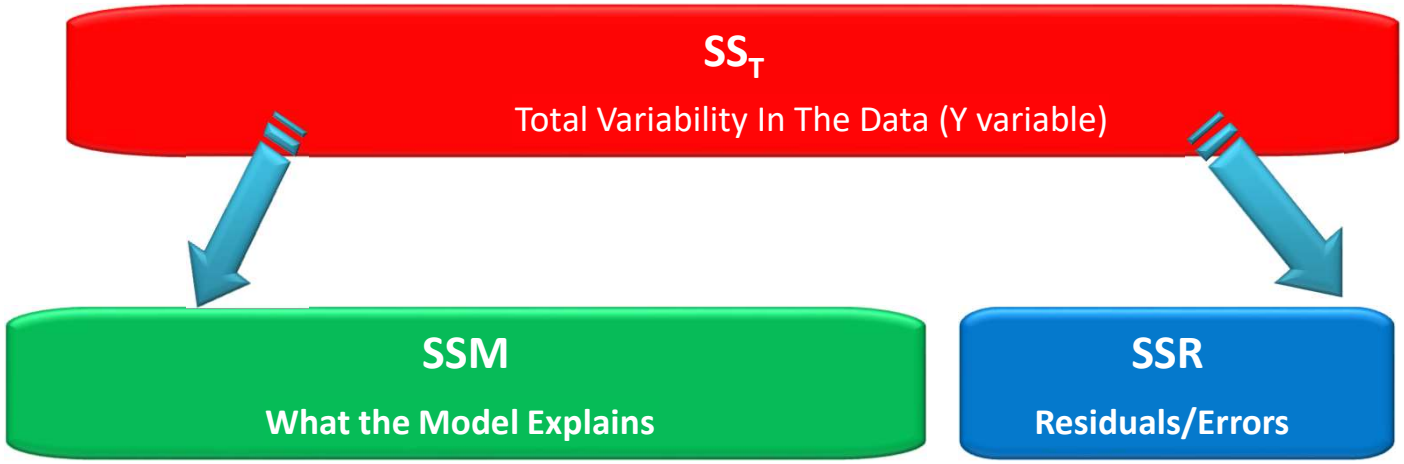
London Business School

Contents

- Further regression examples
- Interaction variables
- Regression Diagnostics

Splitting Variability into *Model* and *Residual*

- SS_T : Total variability between Y variable values and the mean value of Y.
- SS_R : Residual/Error variability (variability between the regression model and the actual data).
- SS_M : Model variability (difference in variability between the model and the mean).

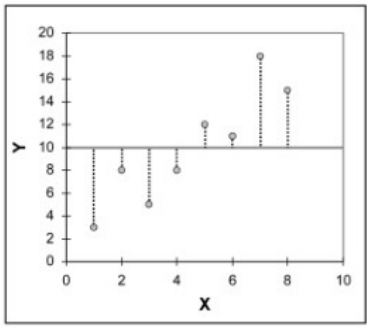


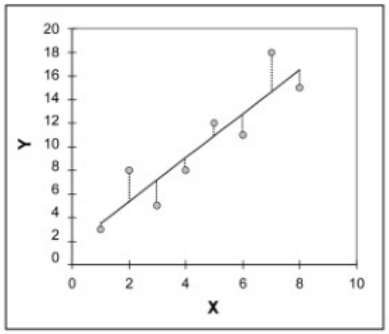
If the regression model results in better prediction than using the mean, then we expect SS_M to be much greater than SS_R

$$R^2 = \frac{SS_M}{SS_T}$$

R^2 is the proportion of the total variability of Y which is explained by the model

Regression: Analysis of Variance (ANOVA)



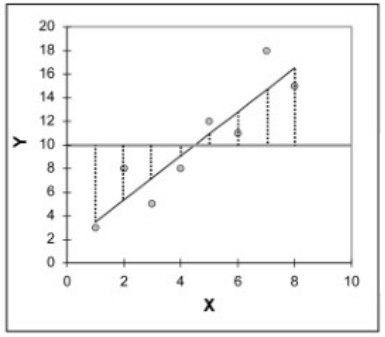


(2)

(3)

SS_T uses the differences between the observed data and the mean value of *Y*

SS_R uses the differences between the observed data and the regression line



(1)

SS_M uses the differences between the mean value of *Y* and the regression line

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   698.93       9.47   73.82  < 2e-16 ***
STR            -2.28       0.48   -4.75  2.8e-06 ***

Residual standard error: 18.6 on 418 degrees of freedom
Multiple R-squared:  0.0512,    Adjusted R-squared:  0.049
F-statistic: 22.6 on 1 and 418 DF,  p-value: 2.78e-06
Analysis of Variance Table

Response: testScore
           Df Sum Sq Mean Sq F value  Pr(>F)
STR         1   7794    7794    22.6 2.8e-06 ***
Residuals 418 144315     345
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
              2.5 % 97.5 %
(Intercept) 680.32 717.54
STR          -3.22  -1.34
```

(1) (2)

TestScore = 698.9 – 2.28* STR + error

- STR (student per teacher ratio) explains a small fraction ($R^2 = 0.0512$) of the variation in test scores.
- Does this mean that STR is not important in a policy sense?

Residual Standard Error

The magnitude of the residuals provide a good indication of how useful the regression line is for predicting *Y* values from *X* values. It is useful to summarize all residuals with a single numerical measure. This measure is called the standard error of estimate and is denoted se. It is essentially the standard deviation of the residuals and is given by this equation:

```
> model1 <- lm(test_score~str,data=CA_test_scores)
> msummary(model1)
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)   698.93       9.47   73.82 < 0.0000000000000002
str            -2.28       0.48   -4.75            0.0000028

Residual standard error: 18.6 on 418 degrees of freedom
Multiple R-squared:  0.0512,    Adjusted R-squared:  0.049
F-statistic: 22.6 on 1 and 418 DF,  p-value: 0.00000278
> anova(model1)
Analysis of Variance Table

Response: test_score
           Df Sum Sq Mean Sq F value    Pr(>F)
str         1   7794    7794    22.6 0.0000028
Residuals 418 144315     345
> confint(model1)
                2.5 %   97.5 %
(Intercept) 680.323 717.5428
str          -3.223  -1.3366
```

The Square Root of MS (Mean Square) is equal to the sum of the squared residuals (SSR) divided by the residuals' degrees of freedom, 418 in this example.

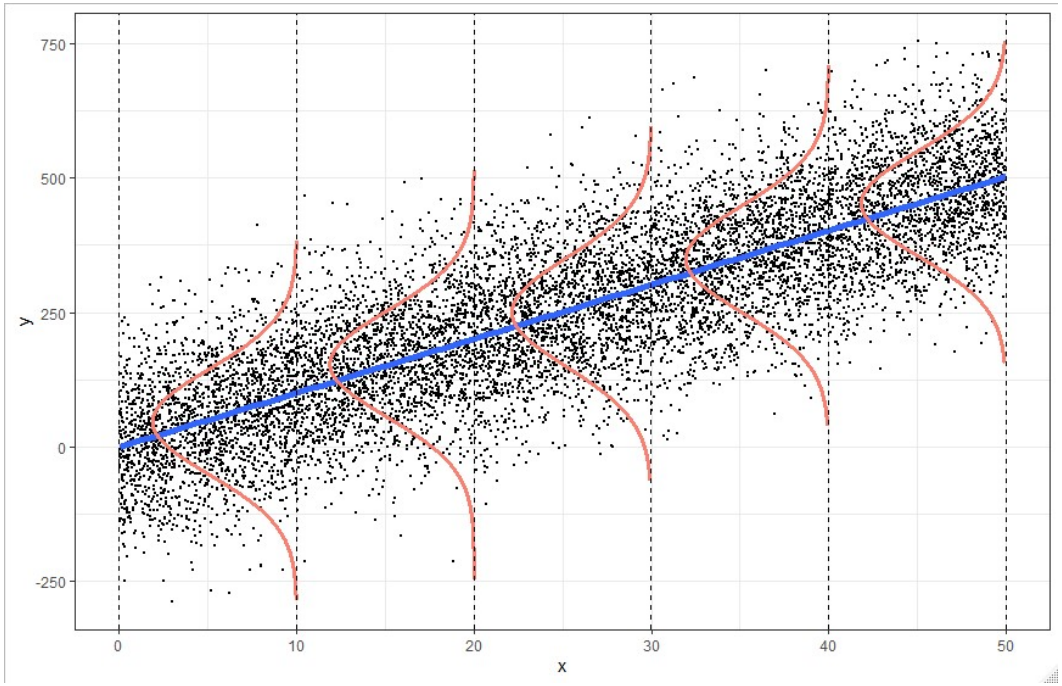
$$Regression\ SE = \sqrt{\frac{114315}{418}} = \sqrt{345} = 18.58$$

In general, the regression standard error indicates the level of accuracy of predictions made from the regression equation.

The smaller it is, the more accurate predictions tend to be.

L-I-N-E: Assumptions of Linear Regression

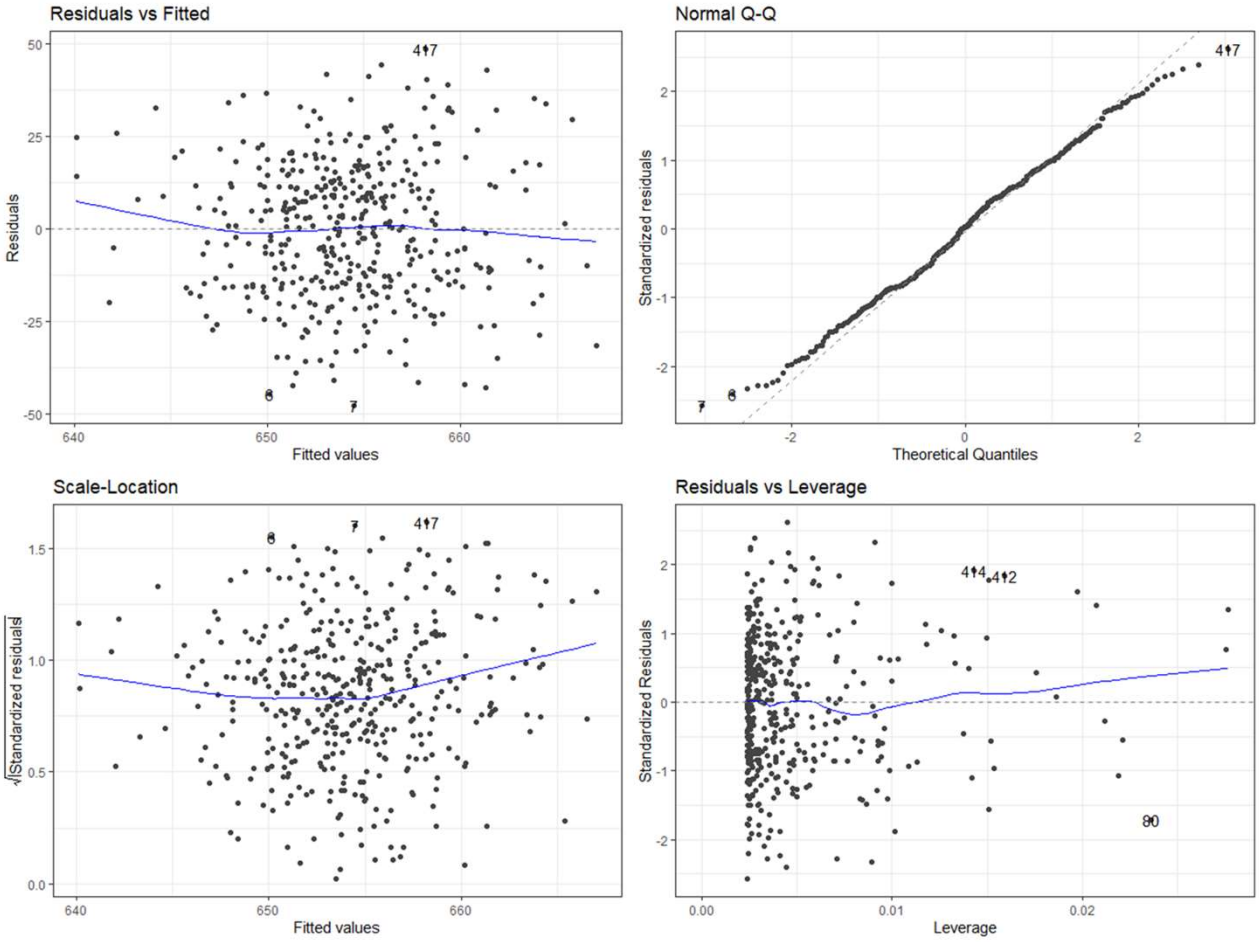
L: Linear relationship between (Y) and the explanatory variable (X)
I: Independence of errors—there's no connection between how far any two points lie from the regression line
N: Normal distribution of Y at each level of X
E: equality of variance of the errors – variability remains the same for all levels of X.



L: The mean value for Y at each level of X lies on regression line.
I: There is no clear pattern in the errors
N: At each level of X, the values for Y are normally distributed.
E: The variability in the Y's for each level of X is the same

Diagnostic Plots for Residuals

1. **Residuals vs. Fitted**: check Linearity assumption. Residuals should be random, with no pattern, and around Y = 0; if not, there is a pattern in the data that is currently unaccounted for.
2. **Normal Q-Q**: check Normality assumption. Deviations from a straight line indicate that residuals do not follow a Normal distribution.
3. **Scale-Location:** check Equal Variance assumption. Positive or negative trends across the fitted values indicate variability that is not constant.
4. **Residuals vs. Leverage**: check for influential points. Points with high leverage (having unusual values of the predictors) and/or high absolute residuals can have an undue influence on estimates of model parameters.



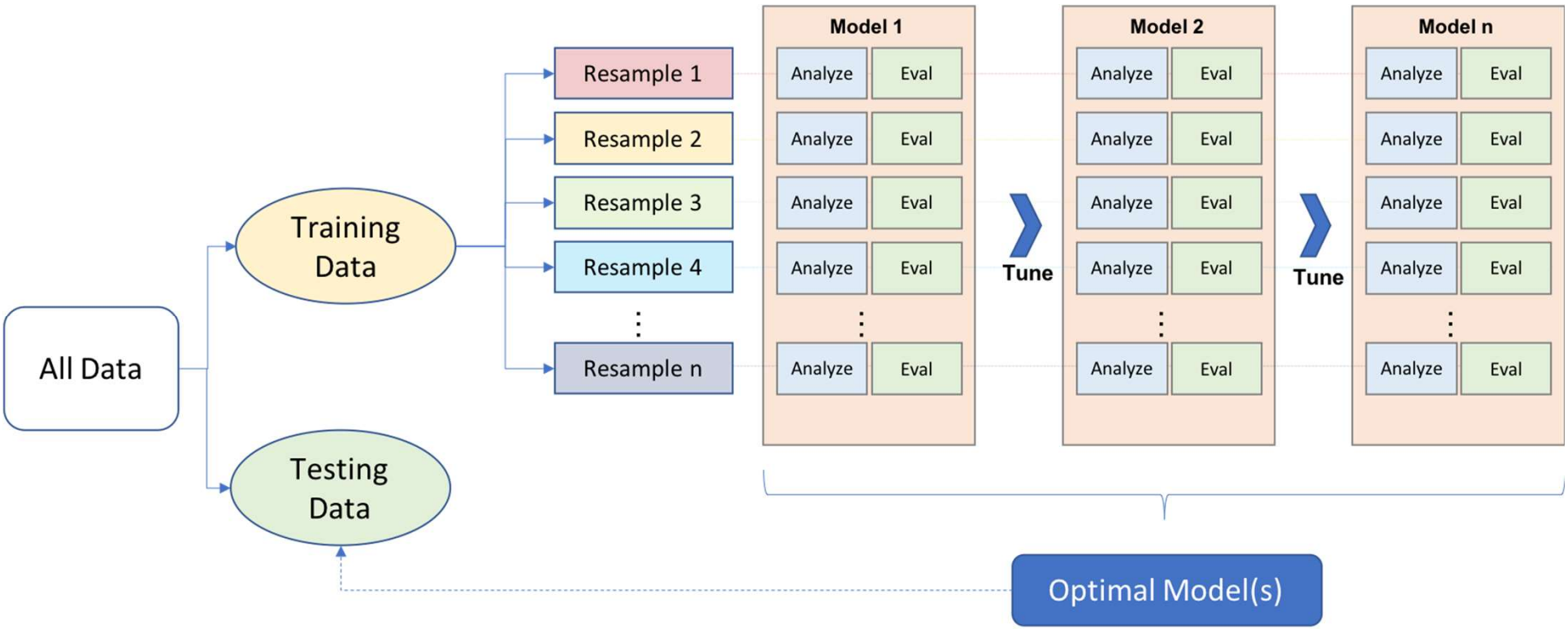

```
# plot residuals
library(ggfortify)
autoplot(model1) +
  theme_bw()
```

Out of sample testing

- Assessing how good a model via adjusted R2 is helpful but could be misleading and lead to overfitting
- Gold standard – out-of-sample goodness of fit
- Hold out method: Randomly partition the data in two sets
  - Training set (~70% of the data): Data in this dataset is used to fit the model
  - Testing (validation) set (~30% of the data): Data in this dataset is used to assess how good is the model
    - Check RMSE (Root Mean Squared Error) and R2 of the validation set and compare them to the training set. If difference is small then overfitting is not a problem. In any case, report the out of sample RMSE and R2

– Typically the fit of out-of-sample is worse than in-sample and more representative of the model's true predictive value -- it does not suffer from overfitting

Out of sample testing



https://bradleyboehmke.github.io/HOML/process.html

London Business School

Splitting data in training and testing sets

Prediction is useful as a way to test the accuracy of your model: split your data into a training set (used for estimation) and a testing set (used for the pseudo-prediction) and see if your model overfits the data.

RMSE, or Root Mean Squared Error is the statistic to calculate relative performance

```
# Splitting data in training and testing sets
library(rsample)
set.seed(1234)  # for reproducibility, and to always get the same split, set the random seed first

train_test_split <- initial_split(CA_test_scores, prop = 0.75)
CA_scores_train <- training(train_test_split)
CA_scores_test <- testing(train_test_split)

rmse_train <- CA_scores_train %>%
  mutate(predictions = predict(model3, .)) %>%
  summarise(sqrt(sum(predictions - test_score)**2/n())) %>%
  pull()
rmse_train
# [1] 2.3212

rmse_test <- CA_scores_test %>%
  mutate(predictions = predict(model3, .)) %>%
  summarise(sqrt(sum(predictions - test_score)**2/n())) %>%
  pull()
rmse_test
# [1] 4.0204
```

London Business School

Workflow: Regression for Prediction

- Clean and inspect the data
- EDA: Exploratory Data Analysis
- Fit several reasonable models using Ordinary Least Squares (OLS) estimation
- Feature engineering: Categorical variables, non-linear terms, interaction variables
- In-sample vs. Out-of-sample testing; check RMSE of train vs. testing dataset.
- Choose a model and use it responsibly!

Regression formula notation in R

Symbol	**Example**	**Meaning**
+	+ X	include this variable in your regression model
-	- X	delete this variable
:	X:Z	include the interaction (***x*z***) between these variables
*	X*Y	include variables and interactions (X, Y, X:Y)
|	X | Z	conditioning: include X given Z
^	(X + Z + W)^3	include variables and all interactions up to three way
I	I (X*Z)	*as is*: include a variable equal to variables multiplied
.		include all explanatory variables in the data frame

London Business School

Beyond linear regression

R has a lot of other built-in functions for regression, such as ***glm***() (for Generalized Linear Models) that we use for binary classification. There are packages for time series, panel data, machine learning, Bayesian and nonparametric methods.

Model	Package	Example
Robust Linear Regression	`MASS`	`rlm(y ~ x, data = mydata)`
Logit	`stats`	`glm(y ~ x, data = mydata, family = "binomial")`
K-means	`stats`	`kmeans(data, n)`
Principal Component Analysis (PCA)	`stats`	`prcomp(data, scale = TRUE, center = TRUE)`
Cox proportional hazards model	`survival`	`coxph(Surv(y_time, y_status) ~ x, data = mydata)`
Panel Data	`plm`	`plm(y ~ x, data = mydata, model = "within|random")`
Time Series	`forecast`	

To do

1. Read ModernDive (www.moderndive.com) , chapters 10.1-10.3, 11

2. Start working on final group assignment.