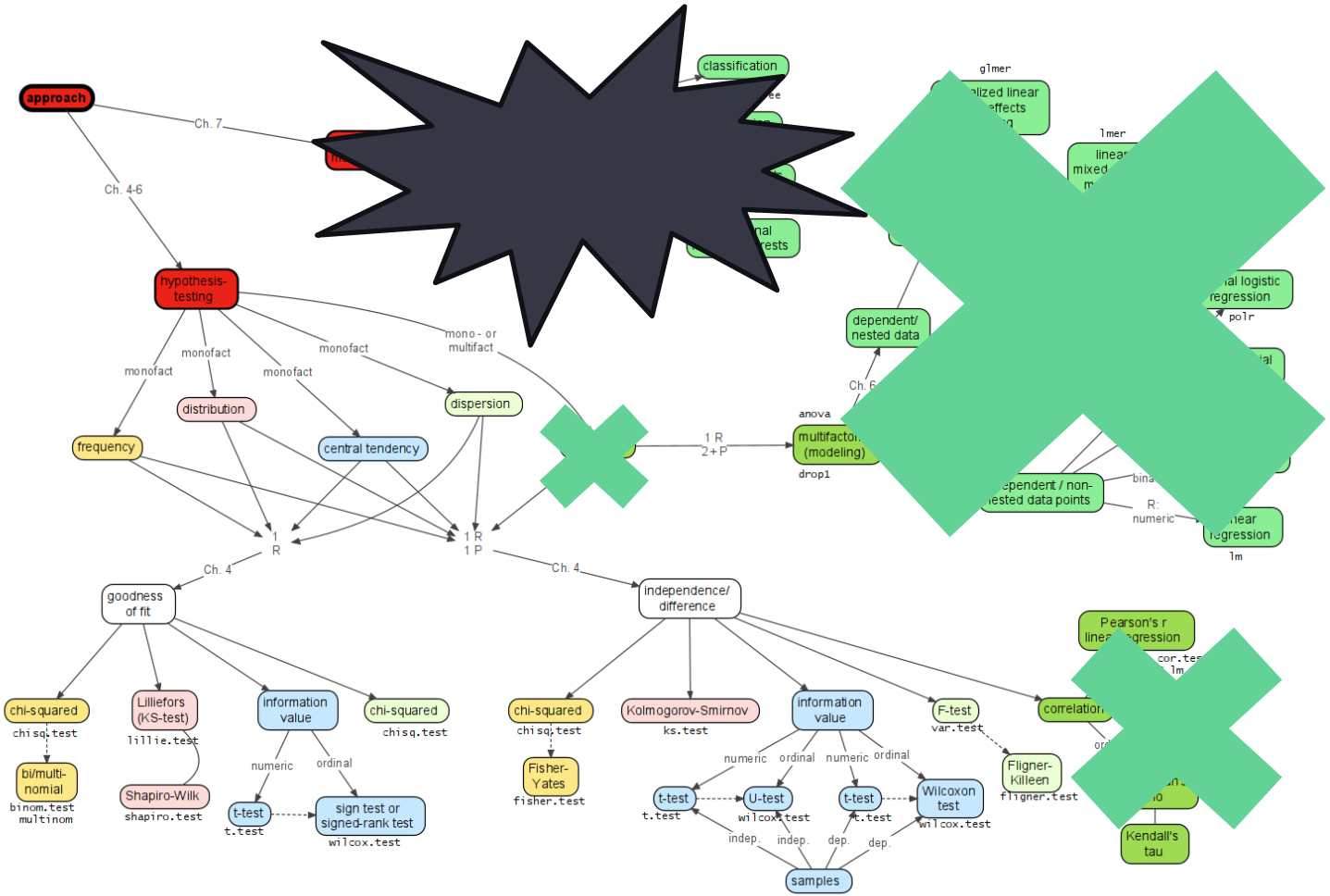
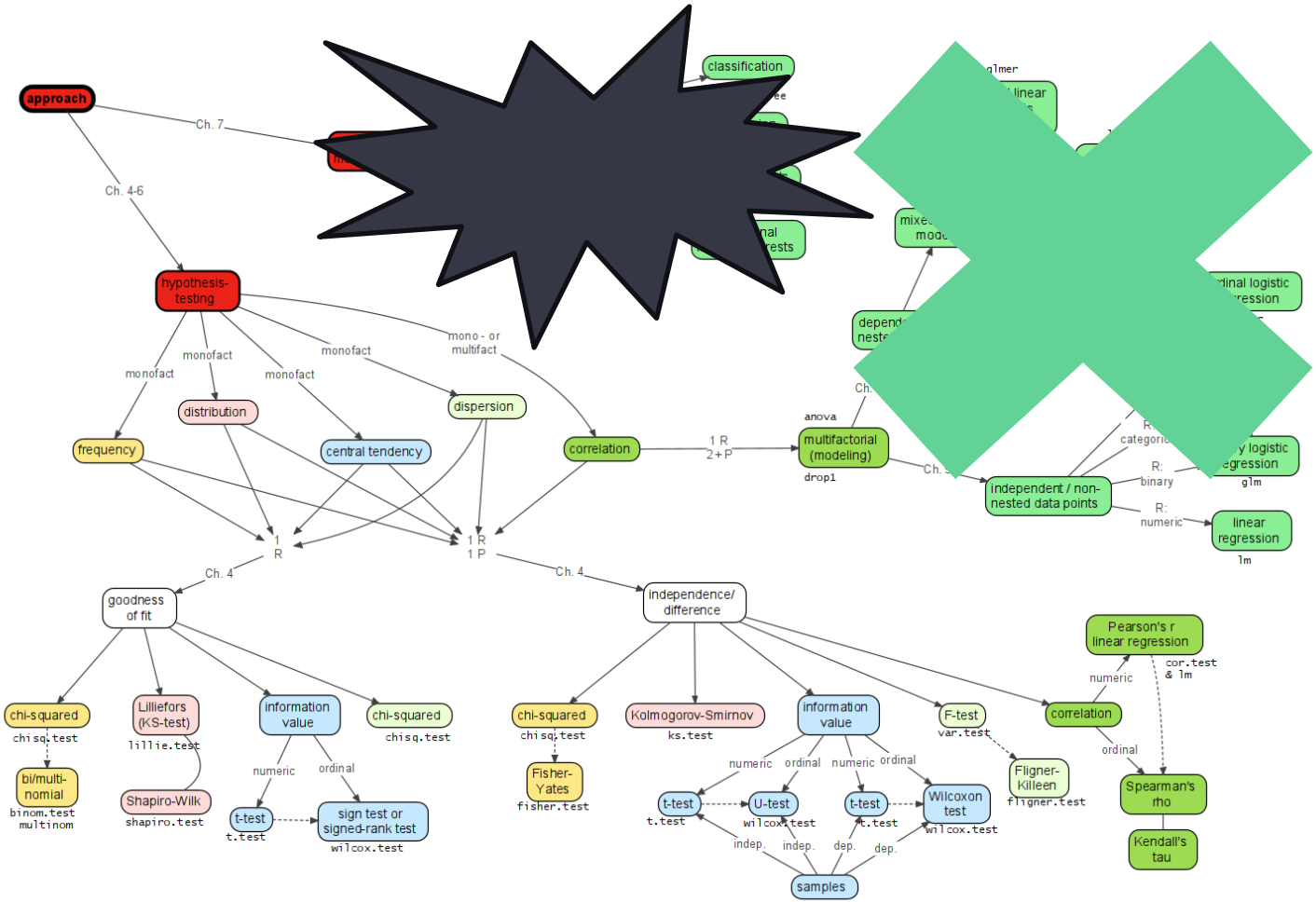
Correlation and linear models

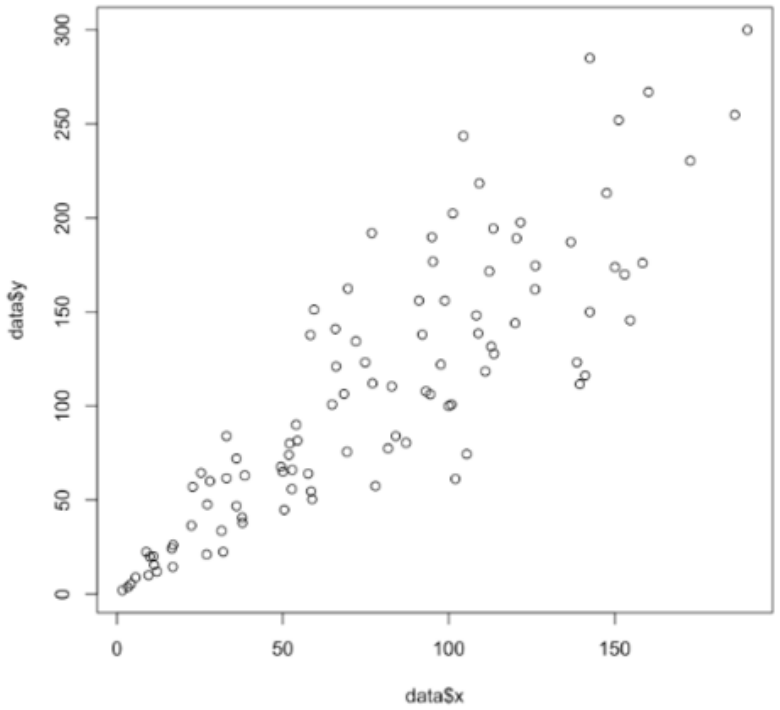

approach
Ch. 7
Ch. 4-6
classification
hypothesis-testing
monofact
monofact
monofact
monofact
mono - or multifact
distribution
dispersion
frequency
central tendency
1 R
1 R 1 P
Ch. 4
goodness of fit
chi-squared
chisq.test
bi/multi-nomial
binom.test multinom
Lilliefors (KS-test)
lillie.test
Shapiro-Wilk
shapiro.test
information value
numeric
ordinal
t-test
t.test
sign test or signed-rank test
wilcox.test
chi-squared
chisq.test
1 R 2+P
anova
multifactor (modeling)
drop1
dependent/ nested data
Ch. 6
glmer
lmer
polr
numeric
lm
Ch. 4
independence/ difference
chi-squared
chisq.test
Fisher-Yates
fisher.test
Kolmogorov-Smirnov
ks.test
information value
numeric
ordinal
numeric
ordinal
t-test
t.test
U-test
wilcox.test
t-test
t.test
Wilcoxon test
wilcox.test
indep.
indep.
dep.
dep.
samples
F-test
var.test
Fligner-Killeen
fligner.test
Pearson's r
cor.test
lm
Kendall's tau


approach
Ch. 7
classification
Ch. 4-6
hypothesis-testing
monofact
monofact
monofact
monofact
mono - or multifact
distribution
dispersion
frequency
central tendency
correlation
1 R
2+P
multifactorial (modeling)
anova
drop1
independent / non-nested data points
R: binary
glm
R: numeric
linear regression
lm
1 R
1 P
Ch. 4
goodness of fit
chi-squared
chisq.test
bi/multi-nomial
binom.test
multinom
Lilliefors (KS-test)
lillie.test
Shapiro-Wilk
shapiro.test
information value
numeric
ordinal
t-test
t.test
sign test or signed-rank test
wilcox.test
chi-squared
chisq.test
Ch. 4
independence/ difference
chi-squared
chisq.test
Fisher-Yates
fisher.test
Kolmogorov-Smirnov
ks.test
information value
numeric
ordinal
t-test
t.test
U-test
wilcox.test
indep.
samples
numeric
ordinal
t-test
t.test
Wilcoxon test
wilcox.test
dep.
F-test
var.test
Fligner-Killeen
fligner.test
correlation
Pearson's r linear regression
cor.test & lm
numeric
ordinal
Spearman's rho
Kendall's tau

Overview

1. **Correlation**
2. Linear model
3. (Understanding predictors)
4. (Model selection)

Steps to determine correlation



1. Visual inspection
    - Is there an association?
    - Linear ➔ regression line
    - Positive or negative? ➔ direction
    - Strength?
- Calculate Pearson's correlation coefficient
    - Between -1 and 1 + p-value
    - sp.stats.pearsonr()

Interpreting the results:

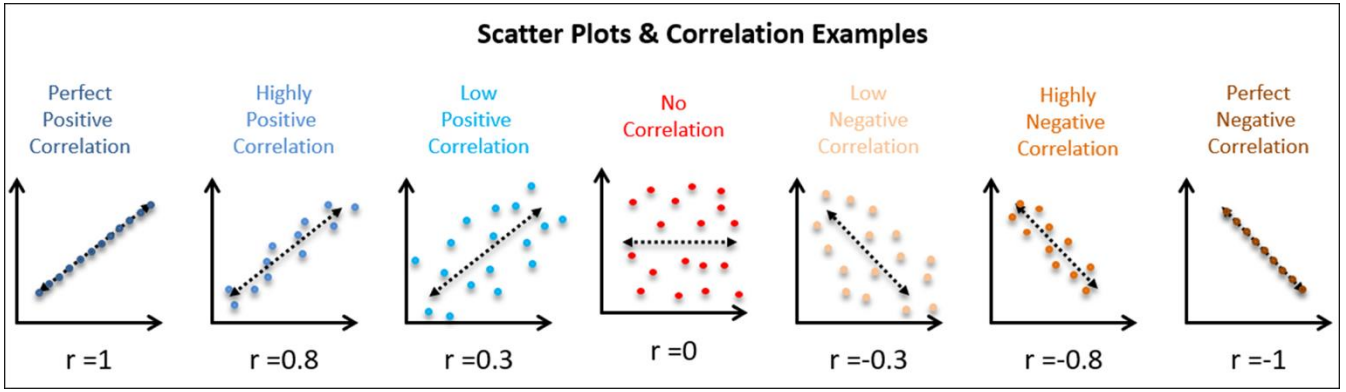


Table 18. Correlation coefficients and their interpretation

Correlation coefficient	Labeling the correlation	Kind of correlation
$0.7 < r \leq 1$	very high	positive correlation:
$0.5 < r \leq 0.7$	high	the more/higher …, the more/higher …
$0.2 < r \leq 0.5$	intermediate	the less/lower …, the less/lower …
$0 < r \leq 0.2$	low	
$r \approx 0$	no statistical correlation (H_0)	
$0 > r \geq -0.2$	low	negative correlation:
$-0.2 > r \geq -0.5$	intermediate	the more/higher …, the less/lower …
$-0.5 > r \geq -0.7$	high	the less/lower …, the more/higher …
$-0.7 > r \geq -1$	very high	

Kendall’s τ & Spearman’s ρ (rank correlation)

- In case both variables are not normally distributed (⇔ discussion)
- Non-parametric alternatives: Kendall’s τ or Spearman’s ρ
    - Similar interpretation: between -1 and 1
    - sp.stats.kendalltau() and sp.stats.spearmanr()
- Time series with Kendall’s τ
    - Has variable become larger/smaller over time?
    - How consistent was the increase/decrease?
    - Time as second variable

When to use (rank) correlation?

- Rank in case of:
  - Dubious interval/ratio variables
  - Less susceptible to outliers
  - Really skewed distributions
  - Small sample sizes
  - Curvilinear distributions

Or
Dependent on level of measurement (always use the lowest!)

When to use (rank) correlation?

Level of measurement	Visual inspection	Test	Measure of effect
Nominal	Contingency table	Chi square	Cramer's V
Ordinal	Contingency table/ scatterplot	Rank correlation	Spearman's ρ Kendall's τ
Interval/ratio	Scatterplot	Correlation	Pearson's r

- Everything is dependent on p-values of course!

Overview

1. Correlation
2. **Linear model**
3. **(Understanding predictors)**
4. **(Model selection)**

Bivariate linear regression

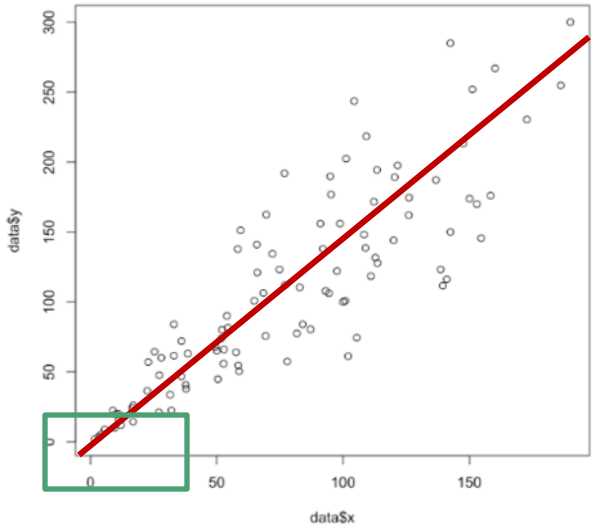
- Predictive/explanatory analysis of two (interval/ratio) variables
- What happens to y if *something* happens to x?
    - Causal association between independent variable x and dependent variable y ⇔ correlation
- Hypotheses:
    - H0: When independent variable x changes, dependent variable y does not change
    - H1: With an increase of 1 in the independent variable x, the dependent variable y increases or decreases
    - = how x **predicts** y
- Example: years of education and income at 30

Other regressions (that we will talk about today…)

- Multiple linear regression
  - Multiple independent variables/predictors that are the same level of measurement
- Multivariate linear regression
  - Multiple independent variables/predictors that are **not** the same level of measurement

Steps to determine and evaluating regression model

1. Visual inspection
    - Scatterplot with regression line
    - Assumption: distribution/dispersion of observations approximates a straight line
2. Calculate formula:
    - y = mx + c + residuals
    - m = slope (regression coefficient)
    - c = intercept/constant

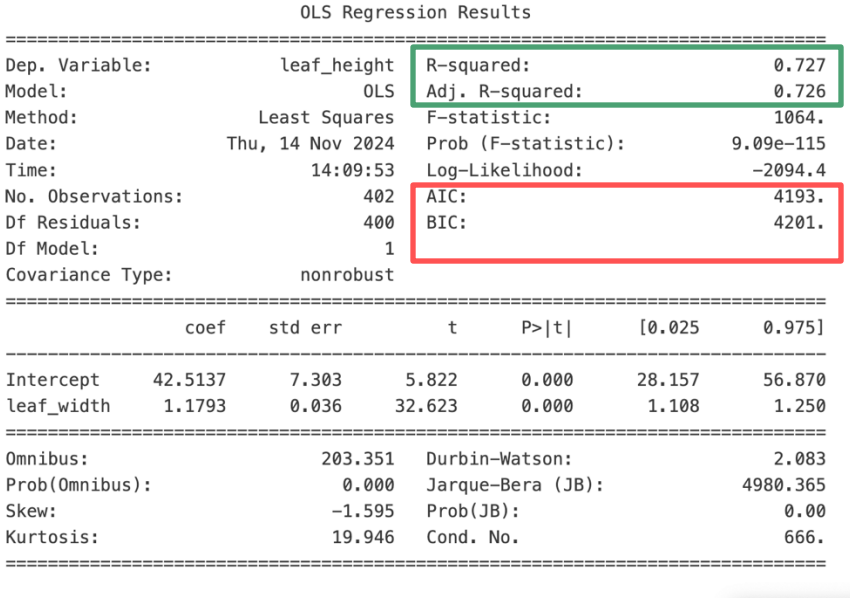


Steps to determining and evaluating regression models

3. Calculate R-squared (coefficient of determination)
    - = squared Pearson's r
    - Between 0 and 1
    - Expresses how much of the total variation in the dependent variable (y) can be explained by variation in the independent variable (x) (= percentage of regression model that predicts dependent variable)
    - Example years of education and income at 30, R^2 = 0,71
        - 71% of the differences in income are explained by education, 29% by other things
    - ≠ significance!

Steps to determining and evaluating regression models

3. Calculate R-squared (coefficient of determination)
    - Also reported: adjusted R^2
    - Used for regressions with multiple variables
        - More is better but watch out for overfitting!
    - Preferred statistic for comparing fits across different models, same interpretation as r^
        - In combination with AIC & BIC (the lower the better)



OLS Regression Results

Dep. Variable:	leaf_height	R-squared:	0.727
Model:	OLS	Adj. R-squared:	0.726
Method:	Least Squares	F-statistic:	1064.
Date:	Thu, 14 Nov 2024	Prob (F-statistic):	9.09e-115
Time:	14:09:53	Log-Likelihood:	-2094.4
No. Observations:	402	AIC:	4193.
Df Residuals:	400	BIC:	4201.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	42.5137	7.303	5.822	0.000	28.157	56.870
leaf_width	1.1793	0.036	32.623	0.000	1.108	1.250

Omnibus:	203.351	Durbin-Watson:	2.083
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4980.365
Skew:	-1.595	Prob(JB):	0.00
Kurtosis:	19.946	Cond. No.	666.

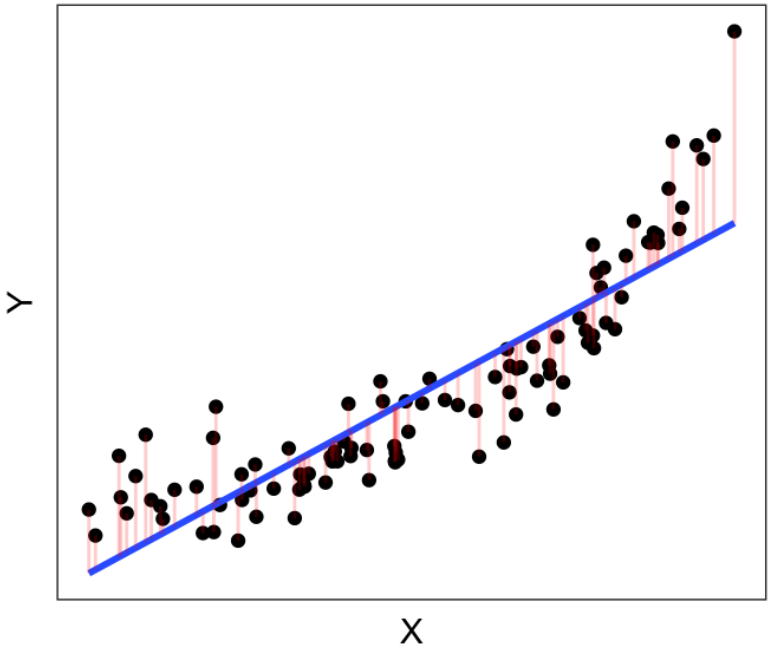
Steps to determining and evaluating regression models

4. Look at model parameters

   - Intercept
   - Predictor/slope e.g. leaf_width
   - Actual estimates & standard error of those estimates
   - T-value & p-value: H0 is that actual intercept and coefficient are 0, H1 that they are 'real'
   - Confidence interval for estimates

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	42.5137	7.303	5.822	0.000	28.157	56.870
leaf_width	1.1793	0.036	32.623	0.000	1.108	1.250

Steps to determining and evaluating regression models



5. Look at residuals

- Should be as minimal as possible
  - And squared so we can interpret them easily
- Should be normally distributed
  - Otherwise, standard error, confidence interval, p-values are not accurate
- Method: bottom Ordinary Least Squares (OLS)
  - Visualises prediction error (= how far the predicted point is from the actual point)

Ok, I can do a regression, now what?

Table 3. *Occupational status over time*

	1750–1	*1760–1*	*1770–1*	*1780–1*	*1790–1*	*1800–1*	*1810–11*	*1820–1*
No. of thefts	51	15	123	75	94	127	155	140
No. where occupational status known	23	7	45	31	52	73	101	69
% in each category:								
Titled	4	–	11	3	–	3	2	3
Professional	4	57	13	10	6	1	3	16
Paperwork	4	–	4	–	4	7	3	4
Non-food seller	39	14	24	10	23	21	35	10
Jeweller	–	14	7	7	4	8	9	9
Innkeeper	–	–	13	13	14	11	12	10
Food seller	9	14	4	7	4	10	10	10
Clothing maker	13	–	9	3	14	15	9	9
Trades	17	–	4	19	8	7	3	10
Army/servant	9	–	2	16	12	1	6	3
Casual	–	–	7	13	14	16	9	16
Mean status	50.2	70.0	58.1	38.9	42.9	45.1	50.9	46.7
% female occupations	4.3	14.3	13.3	9.7	11.5	6.8	5.9	8.6

Notes:

Correlation coefficients:

Pearson's R	–0.063 (0.21 significance)
Spearman's rank	–0.046 (0.36 significance)

Chi-squared:

	1750–81	1790–1821
Jeweller and above	50.9%	43.7%
Innkeeper and below	49.1%	56.3%
(sample)	(106)	(265)
χ^2	1.636 (0.20 significance)	

Regression:

Status rank =	52.658	–0.782 time trend
	(14.15)*	(–1.25)

Adjusted $R^2 = 0.001$, $F = 1.567$, $n = 401$

t-ratio in parentheses.

* Significant at 10% level or higher.

Source: As for tab. 1.

Horrel, Humphries and Sneath, 'Consumption conundrums unravelled', *The Economic History Review* 68:3 (2015), 830-857