Statistiek II

John Nerbonne

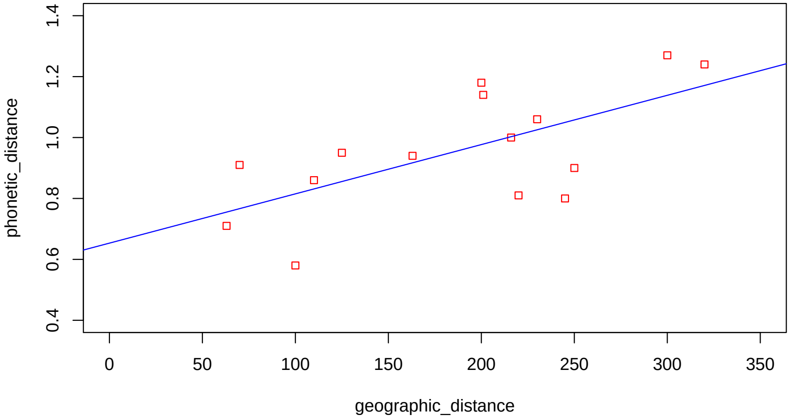
Dept of Information Science
j.nerbonne@rug.nl
incl. important reworkings by Harmut Fitz

March 17, 2015

Review: regression

- compares result on two distinct tests, e.g., geographic and phonetic distance of dialects
- regression for numerical variables only
- fits a straight line on the data
- is there an explanatory relationship between these variables?
- answer: hypothesis tests for regression coefficients
- regression is asymmetric (explanatory direction)
- regression fallacy: seeing causation in regression
- regression towards the mean (inevitable)

Review: regression



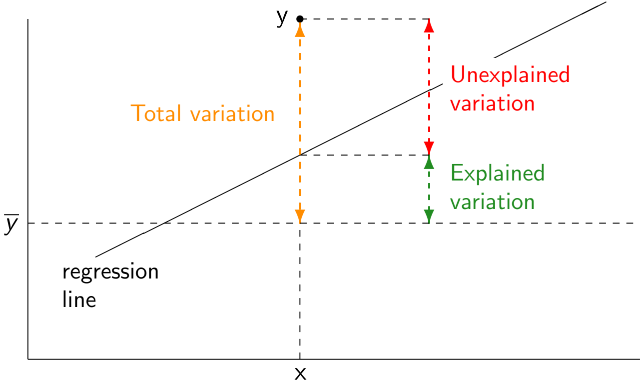
Regression line $y = a + bx$ minimizes the sum of squared residuals

Review: correlation

- only for numeric variables x and y
- measures strength and direction of a linear relation between x and y
- $r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} \cdot z_{y_i}$
- correlation coefficient symmetric: $r_{xy} = r_{yx}$
- $-1 \leq r_{xy} \leq 1$ pure number, no scale
- related to the slope of the regression line: $y = a + bx$ has slope

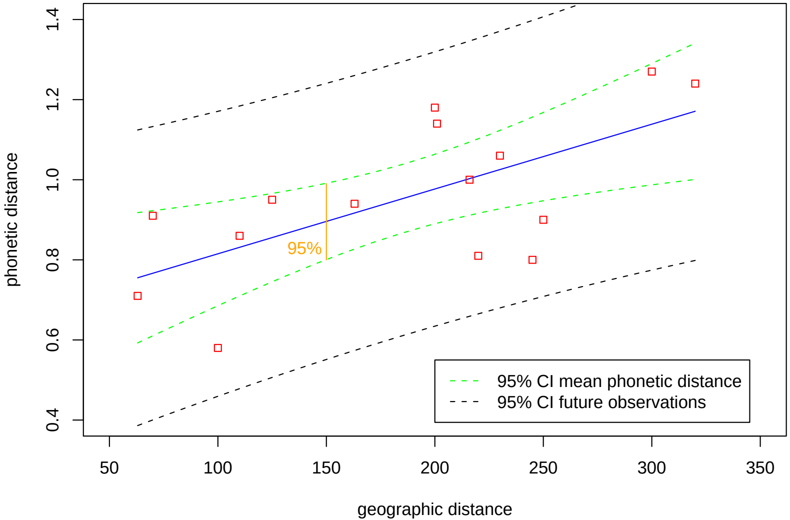
$$b = r \cdot \frac{\sigma_y}{\sigma_x}$$

Review: correlation and regression



Coefficient of determination: $r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$

Review: prediction with regression



Today: multiple regression

Idea: Predict numerical variable using several independent variables

Examples:

- university performance dependent on general intelligence, high school grades, education of parents,...
- income dependent on years of schooling, school performance, general intelligence, income of parents,...
- level of language ability of immigrants depending on
  - leisure contact with natives
  - age at immigration
  - employment-related contact with natives
  - professional qualification
  - duration of stay
  - accommodation

Regression techniques attractive

- allows prediction of one variable value based on one **or more** others
- allows an estimation of the importance of various independent factors (cf. ANOVA)

$$
\begin{aligned}
y &= \epsilon \\
y &= \alpha + \epsilon \\
y &= \alpha + \beta_1 x_1 + \epsilon \\
y &= \alpha + \beta_2 x_2 + \epsilon \\
y &= \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon
\end{aligned}
$$

- which independent factors, taken together or separately, explain the dependent variable the best?

Multiple regression data

One dependent variable y, but **several** predictor variables $x_1, \ldots, x_p$

N cases c_i with $i \in \{1, \ldots, N\}$

Each case c_i has the form $c_i = (x_{i1}, \ldots, x_{ip}, y_i)$

Data:

$$\begin{array}{ll} \text{Case 1:} & c_1 = (x_{11}, \ldots, x_{1p}, y_1) \\ \text{Case 2:} & c_2 = (x_{21}, \ldots, x_{2p}, y_2) \\ \vdots & \vdots \\ \text{Case N:} & c_N = (x_{N1}, \ldots, x_{Np}, y_N) \end{array}$$

Example: do geographic (x_1) and phonetic distance (x_2) predict people's intuitions about dialect distance (y)? (see Bezooijen and Heeringa, 2006)

Multiple regression model

Statistical **model** of multiple linear regression:

$$
\begin{aligned}
y_1 &= \alpha + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + \epsilon_1 \\
&\vdots \\
y_N &= \alpha + \beta_1 x_{N1} + \beta_2 x_{N2} + \ldots + \beta_p x_{Np} + \epsilon_N
\end{aligned}
$$

Mean response μ_y is linear combination of predictor variables:

$$\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Deviations ϵ_i are independent and normally distributed with mean 0 and standard deviation σ

Multiple regression model

Need to **estimate** $p+1$ model parameters $a, b_1, \ldots, b_p$:

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

Multiple regression model

Need to **estimate** $p+1$ model parameters $a, b_1, \ldots, b_p$:

$$\underbrace{y = a + b_1 x_1}_{\text{simple linear regression}} + b_2 x_2 + \cdots + b_p x_p$$

Multiple regression model

Need to **estimate** $p+1$ model parameters $a, b_1, \ldots, b_p$:

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

Predicted response for case i:

$$\hat{y}_i = a + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$$

Residual of case i:

$$\begin{aligned} e_i &= \text{observed response} - \text{predicted response} \\ &= y_i - \hat{y}_i \\ &= y_i - a - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_p x_{ip} \end{aligned}$$

Least squares regression

Find parameters that minimize sum of squared residuals (SSE):

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - a - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_p x_{ip})^2$$

But this time, let software do it for you...

As usual, we partition the variance:

$$\begin{aligned} \text{SST} &= \text{SSM} + \text{SSE} \\ \sum_{i=1}^{N} (y_i - \overline{y})^2 &= \sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \\ \text{Total variance} &= \text{Explained variance} + \text{Error variance} \end{aligned}$$

Degrees of freedom in multiple regression

Multiple linear regression model has $p + 1$ parameters

Hence, **model** degrees of freedom (DFM): $(p + 1) - 1 = p$

Total degrees of freedom (DFT): (number of cases) $-1 = N - 1$

Error degrees of freedom (DFE): $N - p - 1$

As usual, DFT = DFM + DFE

Mean square model: MSM = SSM/DFM

Mean square error: MSE = SSE/DFE

Multiple regression: example

Grade point average (GPA) of first-year computer science majors is measured (A $=$ 4.0, B $=$ 3.0,...)

Questions:

(a) do high school grades predict university grades?

- Mathematics
- English
- Science

(b) do ‘scholastic aptitude test’ (SAT) scores predict university grades?

- Mathematics
- Verbal

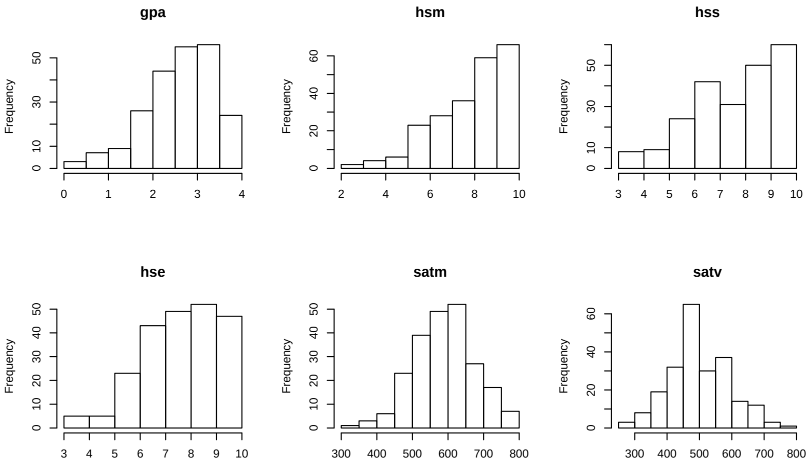
(c) do both sets of scores predict GPA?

Multiple regression: example

Obs	HS-M	HS-S	HS-E	SAT-M	SAT-V	GPA
1	10	10	10	670	600	3.32
2	6	8	5	700	640	2.26
3	8	6	8	640	530	2.35
4	9	10	7	670	600	2.08
5	8	9	8	540	580	3.38
⋮	⋮	⋮	⋮	⋮	⋮	⋮
224	9	8	9	559	488	2.28

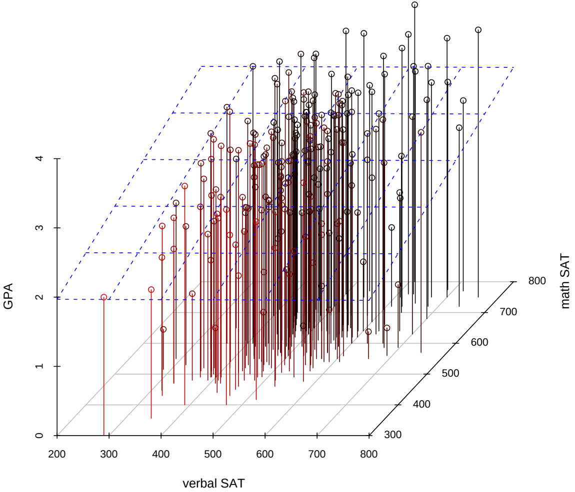
HS-M/S/E: high school grades mathematics/science/English
SAT-M/V: 'scholastic aptitude test' scores mathematics/verbal
GPA: grade point average

Distribution of scores



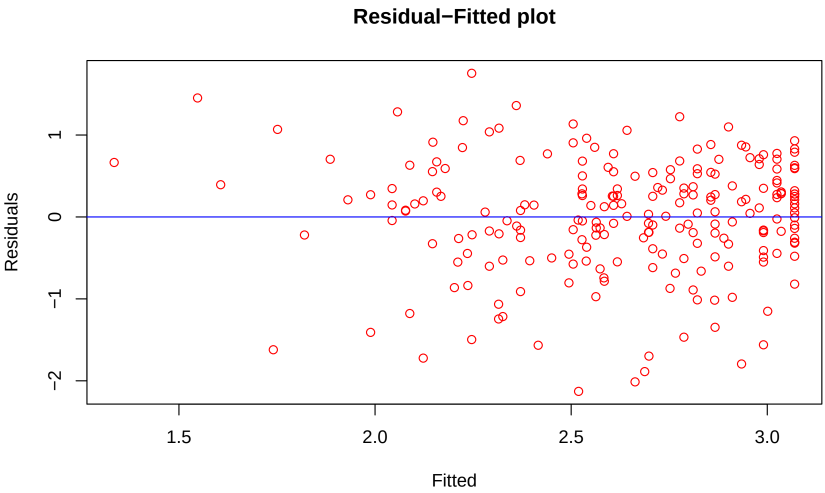
Regression does not require that variables be normally distributed!

Multiple regression: predicted vs observed values



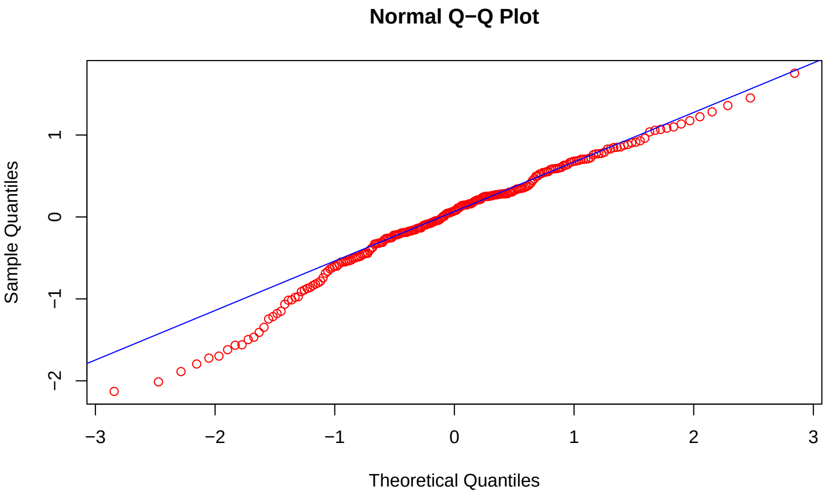
Scatterplot of GPA against SAT scores with regression plane fitted

Visualizing residuals



No indication of non-linear relationship between variables

Check normality of residuals



No indication that residuals are distributed non-normal

Regression on high school grades

(a) do high school grades (HS-M, HS-S, HS-E) predict GPA?

Call: lm(formula = gpa $\sim$ hse + hsm + hss, data = gpa_data)

Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.58988	0.29424	2.005	0.0462 *
hse	0.04510	0.03870	1.166	0.2451
hsm	0.16857	0.03549	4.749	3.68e-06 ***
hss	0.03432	0.03756	0.914	0.3619
—				
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
Residual standard error:	0.6998 on 220 degrees of freedom			
Multiple R-Squared:	0.2046,	Adjusted R-squared: 0.1937		
F-statistic:	18.86 on 3 and 220 DF,		p-value: 6.359e-11	

Regression on high school grades

(a) do high school grades (HS-M, HS-S, HS-E) predict GPA?

Call: $\text{lm}(\text{formula} = \text{gpa} \sim \text{hse} + \text{hsm} + \text{hss}, \text{data} = \text{gpa\_data})$

Coefficients:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.58988	0.29424	2.005	0.0462 *
hse	0.04510	0.03870	1.166	0.2451
hsm	0.16857	0.03549	4.749	3.68e-06 ***
hss	0.03432	0.03756	0.914	0.3619

—

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
Residual standard error:	0.6998 on 220 degrees of freedom			
Multiple R-Squared:	0.2046,	Adjusted R-squared: 0.1937		
F-statistic:	18.86 on 3 and 220 DF,		p-value: 6.359e-11	

Regression equation: $y = 0.59 + 0.04x_1 + 0.17x_2 + 0.03x_3$

Regression on high school grades

(a) do high school grades (HS-M, HS-S, HS-E) predict GPA?

Call: lm(formula = gpa ∼ hse + hsm + hss, data = gpa_data)

Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.58988	0.29424	2.005	0.0462 *
hse	0.04510	0.03870	1.166	0.2451
hsm	0.16857	0.03549	4.749	3.68e-06 ***
hss	0.03432	0.03756	0.914	0.3619
—				
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
Residual standard error:	0.6998 on 220 degrees of freedom			
Multiple R-Squared:	0.2046,	Adjusted R-squared: 0.1937		
F-statistic:	18.86 on 3 and 220 DF,		p-value: 6.359e-11	

Regression equation: $y = 0.59 + 0.04x_1 + 0.17x_2 + 0.03x_3$

F-statistics for multiple regression

F-statistics tests:

H_0: $b_1 = b_2 = \ldots = b_p = 0$ against H_a: at least one of the $b_i \neq 0$

ANOVA table:

Source	Degrees of freedom	Sum of squares	Mean square	F
Model	p	$\sum(\hat{y}_i - \overline{y})^2$	SSM/DFM	MSM/MSE
Error	$N - p - 1$	$\sum(y_i - \hat{y}_i)^2$	SSE/DFE	
Total	$N - 1$	$\sum(y_i - \overline{y})^2$	SST/DFT	

In the example: $F(3, 220) = 18.86$ and $p < 0.001$

Hence, we reject H_0, at least one regression coefficient $b_i \neq 0$ (but we don't know which one)

Regression on high school grades

(a) do high school grades (HS-M, HS-S, HS-E) predict GPA?

Call: lm(formula = gpa ~ hse + hsm + hss, data = gpa_data)

Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.58988	0.29424	2.005	0.0462 *
hse	0.04510	0.03870	1.166	0.2451
hsm	0.16857	0.03549	4.749	3.68e-06 ***
hss	0.03432	0.03756	0.914	0.3619

—

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .

Residual standard error: 0.6998 on 220 degrees of freedom

Multiple R-Squared: 0.2046, Adjusted R-squared: 0.1937

F-statistic: 18.86 on 3 and 220 DF, p-value: 6.359e-11

Regression equation: $y = 0.59 + 0.04x_1 + 0.17x_2 + 0.03x_3$

Hypothesis testing

Which of the high school grades significantly contributes to predicting GPA?

For each coefficient b_1, b_2, b_3 we test: H_0: $b_i = 0$ vs H_a: $b_i \neq 0$

Under H_0: $$t^* = \frac{b_i}{\mathrm{SE}_i}$$

follows t-distribution with $N - p - 1$ degrees of freedom, where

$$\mathrm{SE}_i = \text{standard error of the estimated } b_i$$

If $t^* \geq |t(N - p - 1)|$ at $\alpha = 0.05$, reject H_0

Hypothesis testing

Which of the high school grades significantly contributes to predicting GPA?

Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)
hse	0.04510	0.03870	1.166	0.2451
hsm	0.16857	0.03549	4.749	3.68e-06 ***
hss	0.03432	0.03756	0.914	0.3619

—

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .

In <u>this</u> regression model, only high school grades in Mathematics (HS-M) are significant

BUT...

Hypothesis testing

...if we regress Science grades (HS-S) **only** on GPA:

Call: lm(formula = gpa ∼ hss, data = gpa_data)

Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	1.41325	0.24017	5.884	1.46e-08 ***
hss	0.15106	0.02906	5.198	4.55e-07 ***

—

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
Residual standard error:	0.7375 on 222 degrees of freedom			
Multiple R-Squared:	0.1085,	Adjusted R-squared: 0.1045		
F-statistic:	27.02 on 1 and 222 DF,		p-value: 4.552e-07	

We find that HS-S is a significant predictor of GPA!

Hypothesis testing

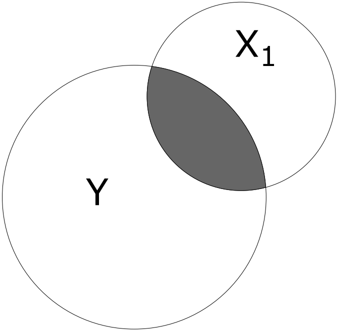
Explanation: look at correlation between explanatory variables

$$\begin{aligned} r_{\text{HSM,HSE}} &= 0.47 \\ r_{\text{HSM,HSS}} &= 0.58 \\ r_{\text{HSE,HSS}} &= 0.58 \end{aligned}$$

Hence, Maths and Science grades strongly correlated

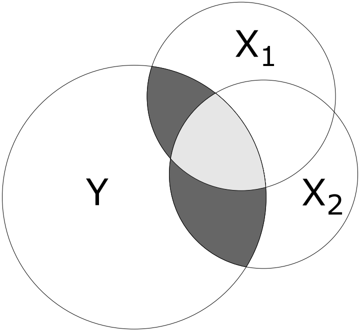
- HSS does not add to explanatory power of HSM and HSE (in full model)
- HSS alone, though, predicts GPA (to some extent)
- be careful: always compare several multiple regression models and determine correlation before drawing conclusions

Visualizing multiple regression



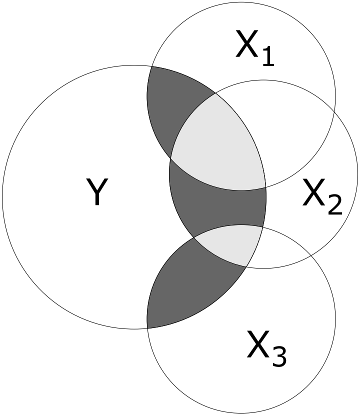
- regress Y on X_1 (simple linear regression)
- shaded area r^2 (squared Pearson correlation coefficient)
- r^2 measures amount of variation in Y explained by X_1

Visualizing multiple regression



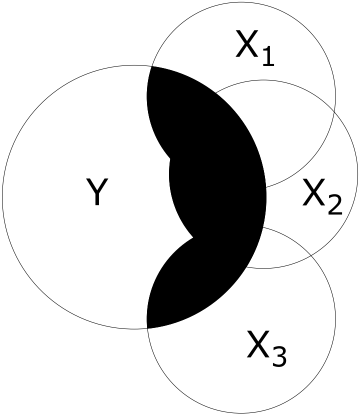
- regress Y on X_1 **and** X_2 (multiple linear regression)
- dark grey areas: **uniquely** explained variance ("squared semi-partial correlation")
- light grey area: **commonly** explained variance (due to correlation of X_1 and X_2)

Visualizing multiple regression



- regress Y on X_1 **and** X_2 **and** X_3 (multiple linear regression)
- dark grey areas: **uniquely** explained variance ("squared semi-partial correlation")
- light grey area: **commonly** explained variance (due to correlation of X_1 and X_2)
- note: X_1 and X_3 uncorrelated

Visualizing multiple regression



- regress Y on X_1 **and** X_2 **and** X_3 (multiple linear regression)
- black area R^2: "squared multiple correlation coefficient"
- R^2 measures total proportion of variance in Y accounted for by X_1, X_2 and X_3

Squared multiple correlation

$$R^2 = \frac{SSM}{SST} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}$$

Regression of GPA on HS-S, HS-M and HS-E:

Residual standard error:	0.6998 on 220 degrees of freedom	
Multiple R-Squared:	0.2046,	Adjusted R-squared: 0.1937
F-statistic:	18.86 on 3 and 220 DF,	p-value: 6.359e-11

- High school grades explain 20.5% of variance in GPA
- Not a whole lot, despite highly significant *p*-value for HS-M coefficient
- Once again, small *p*-values do not entail a large effect!

Squared multiple correlation

$$R^2 = \frac{SSM}{SST} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}$$

Regression of GPA on HS-S only:

Residual standard error:	0.7375 on 222 degrees of freedom	
Multiple R-Squared:	0.1085,	Adjusted R-squared: 0.1045
F-statistic:	27.02 on 1 and 222 DF,	p-value: 4.552e-07

- p-values in both models comparable, but
- High school grades in Science explain only 10.8% of variance in GPA
- Adding more variables (HS-M, HS-E) to model adds explanatory power

Refining the model

In full model (HS-S/E/M), HS-S had largest *p*-value (0.3619); drop HS-S from model:

Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.62423	0.29172	2.140	0.0335 *
hse	0.06067	0.03473	1.747	0.0820 .
hsm	0.18265	0.03196	5.716	3.51e-08 ***

—

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
Residual standard error:	0.6996 on 221 degrees of freedom			
Multiple R-Squared:	0.2016,	Adjusted R-squared: 0.1943		
F-statistic:	27.89 on 2 and 221 DF,		p-value: 1.577e-11	

- $R^2 = 0.2016$ versus $R^2 = 0.2046$ in the bigger model
- In this (precise) sense HS-S does not add to explanatory power

What about SAT scores?

Question (b) do SAT scores predict GPA?

Call: lm(formula = gpa ~ satm + satv, data = gpa_data)

Coefficients:	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	1.289e+00	3.760e-01	3.427	0.000728 ***
satm	2.283e-03	6.629e-04	3.444	0.000687 ***
satv	-2.456e-05	6.185e-04	-0.040	0.968357
—				
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
Residual standard error:	0.7577 on 221 degrees of freedom			
Multiple R-Squared:	0.06337,	Adjusted R-squared: 0.05498		
F-statistic:	7.476 on 2 and 221 DF,		p-value: 0.0007218	

Regression on SAT scores also significant, but less explanatory power than high school grades

What about adding SAT scores?

Question (c) do high school grades **and** SAT scores predict GPA?

Call: lm(formula = gpa ∼ hse + hsm + hss + satm + satv, data = gpa_data)

Coefficients:	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.3267187	0.3999964	0.817	0.414932
hse	0.0552926	0.0395687	1.397	0.163719
hsm	0.1459611	0.0392610	3.718	0.000256 ***
hss	0.0359053	0.0377984	0.950	0.343207
satm	0.0009436	0.0006857	1.376	0.170176
satv	-0.0004078	0.0005919	-0.689	0.491518
—				
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
Residual standard error:	0.7 on 218 degrees of freedom			
Multiple R-Squared:	0.2115,	Adjusted R-squared: 0.1934		
F-statistic:	11.69 on 5 and 218 DF,		p-value: 5.058e-10	

ANOVA for multiple regression

- How do we formally compare different regression models?
- For example, do SAT scores significantly add to explanatory power of high school grades?

Compare
lm(formula = gpa ~ hse + hsm + hss, data = gpa_data)

with
lm(formula = gpa ~ hse + hsm + hss + satm + satv, data = gpa_data)

Use ANOVA to test:

H_0: $b_{satm} = b_{satv} = 0$ versus H_a: at least one of these $b's \neq 0$

ANOVA for multiple regression

ANOVA F-score:

$$F = [(\text{SSE}_{\text{shorter}} - \text{SSE}_{\text{longer}})/\#\text{new variables}]/\text{MSE}_{\text{longer}}$$

In the example:

Analysis of Variance Table

Model 1: gpa $\sim$ hse + hsm + hss
Model 2: gpa $\sim$ hse + hsm + hss + satm + satv

	Res.Df	SSE	Df	Sum of Sq	F	Pr($>F$)
1	220	107.750				
2	218	106.819	2	0.931	0.9503	0.3882

Hence, SAT scores not significant predictors of GPA in regression model which already contains high school scores

Analyses summary

What can we conclude from all these analyses?

- High school grades in Maths are a significant predictor of GPA
- High school grades in Science are a significant predictor of GPA
- High school grades in Science and English do not add to the explanatory power of Math grades
- SAT scores do not add explanatory power to the model either

Can we ignore SAT scores and Science/English grades then?

- No, because we only looked at GPA of computer science majors
- at one university

Problems with multiple regression

- Overfitting: The more variables, the higher the amount of variance you can explain. Even if each variable doesn't explain much, adding large number of variables can result in high values of R^2
- Interaction: Multiple regression is logically more complicated than simple regression applied several times for different variables
- Collinearity: Independent variables may correlate themselves, competing in their explanation
  - Consider "cleaning" one indep. variable of another by using residuals of regression analysis.
- Suppression: An independent variable may appear not to be explanatory, but becomes significant in combined model

Summary multiple regression

- **generalization** of simple linear regression
- allows prediction of one variable value based on one **or more** others
- **test hypotheses** about the predictive power of variables (t-test for coefficients)
- measure the proportion of variance in dependent variable **explained** by predictors (R^2)
- allows an **estimation** of the importance of various independent factors (model comparison with ANOVA)
- which independent factors, taken together or separately, explain the dependent variable the **best**?

Next week

Next week: logistic regression