Lecture 25: Multiple Linear Regression

- Multiple linear regression
- Adjusted R^2
- The Harris trust and savings bank case

Introduction

– In recent lectures we talked about least squares method and simple linear regression.

– We often come across data that do not necessarily satisfy the normal distribution or linearity assumptions for the response variable and we briefly talked about some common transformations that result in normality (or as least as close as possible).

– Today we will extend the regression idea to multiple explanatory variables.

Simple Linear Regression

Recall the simple linear regression assumptions:

1. Each point (x_i, y_i) in the scatterplot satisfies:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where the ϵ_i have a normal distribution with mean zero and (usually) unknown standard deviation.

2. The errors ϵ_i have nothing to do with one another (independence). A large error does not tend to be followed by another large error, for example.
3. The x_i values are measured without error. (Thus all the error occurs in the vertical direction.)
4. The errors ϵ_i are independent of the x_i's.
5. The relationship between y_i and x_i is linear.

Multiple regression

- Simple linear regression: Bivariate - two variables: y and x
- Multiple linear regression: Multiple variables: y and $x_1, x_2, \cdots$

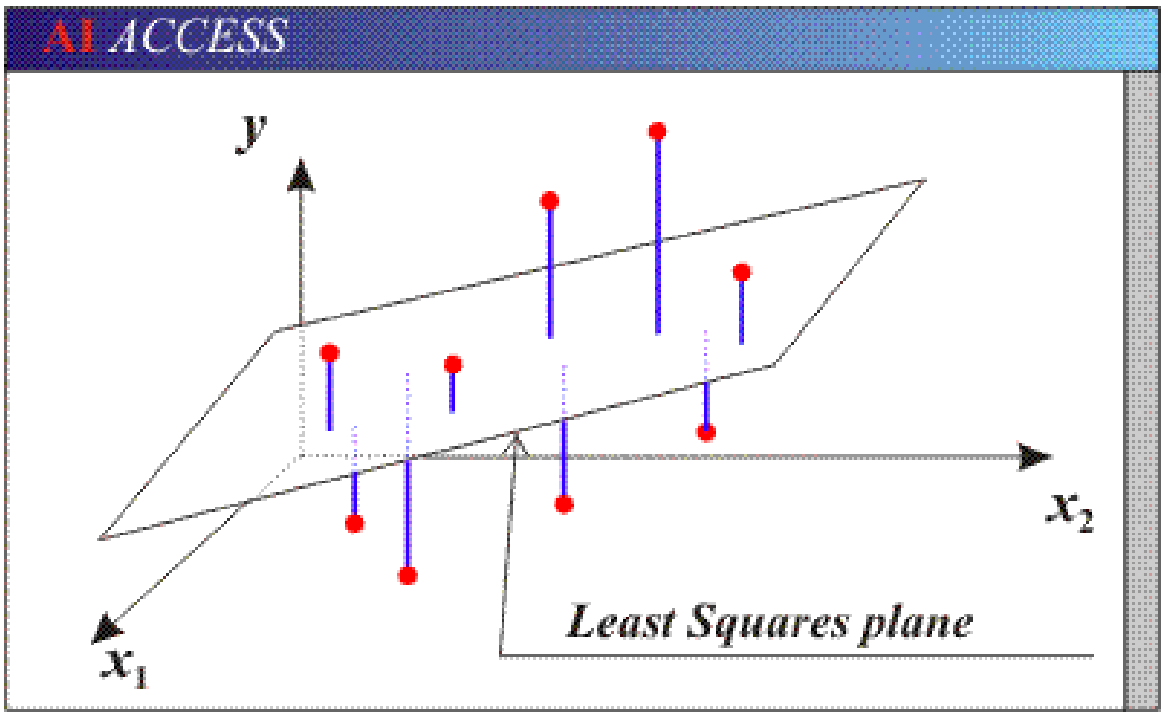
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

Multiple Regression

In multiple regression, the set-up is still the same except that we now have more than one explanatory variable. The model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_p x_{pi} + \epsilon_i.$$

Again, the ϵ_i are independent normal random variables with mean 0 and all the x's are measured without error.



Multiple Regression

How then should we interpret the coefficients (that is, $\beta_0, \beta_1, \ldots, \beta_p$)?

Recall that for the simple linear regression, β_0 is usually the average of y when x is zero (and it is only meaningful when x can be zero) while for β_1, every unit increase in x corresponds to an increase in y by β_1.

The interpretation is quite similar here too except that for any β_p, every unit increase in x_p corresponds to an increase in y by β_p **when all the other x's are fixed or held constant.**

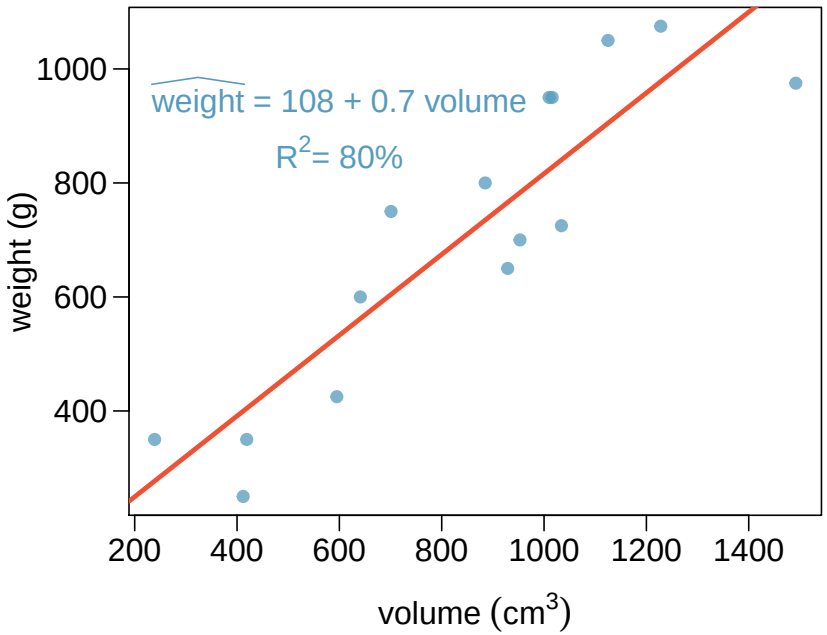
Weights of books

	weight (g)	volume (cm^3)	cover
1	800	885	hc
2	950	1016	hc
3	1050	1125	hc
4	350	239	hc
5	750	701	hc
6	600	641	hc
7	1075	1228	hc
8	250	412	pb
9	700	953	pb
10	650	929	pb
11	975	1492	pb
12	350	419	pb
13	950	1010	pb
14	425	595	pb
15	725	1034	pb

From: Maindonald, J.H. and Braun, W.J. (2nd ed., 2007) “Data Analysis and Graphics Using R"

Weights of books (cont.)

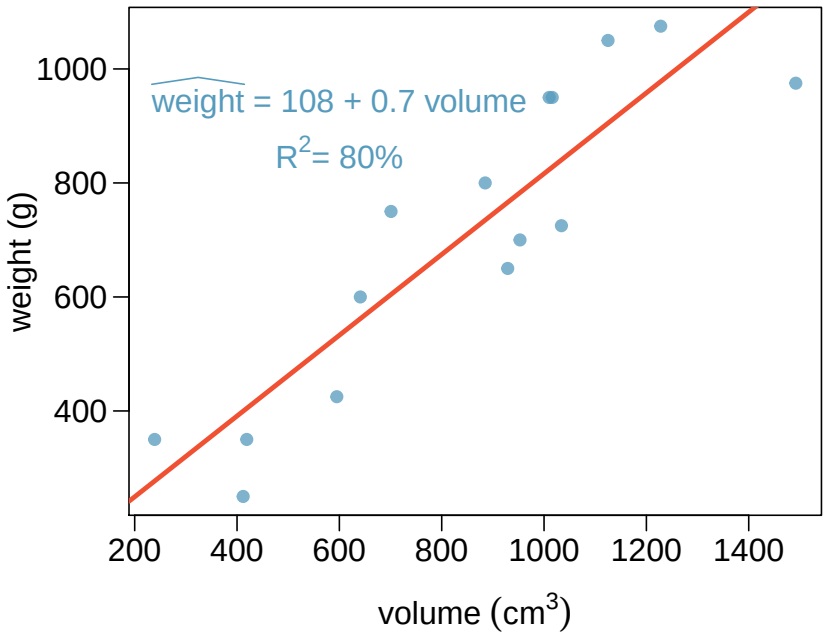
The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



(a) Weights of 80% of the books can be predicted accurately using this model.
(b) Books that are 10 cm^3 over average are expected to weigh 7 g over average.
(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.
(d) The model underestimates the weight of the book with the highest volume.

Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



(a) Weights of 80% of the books can be predicted accurately using this model.

(b) *Books that are 10 cm^3 over average are expected to weigh 7 g over average.*

(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

(d) The model underestimates the weight of the book with the highest volume.

Modeling weights of books using volume

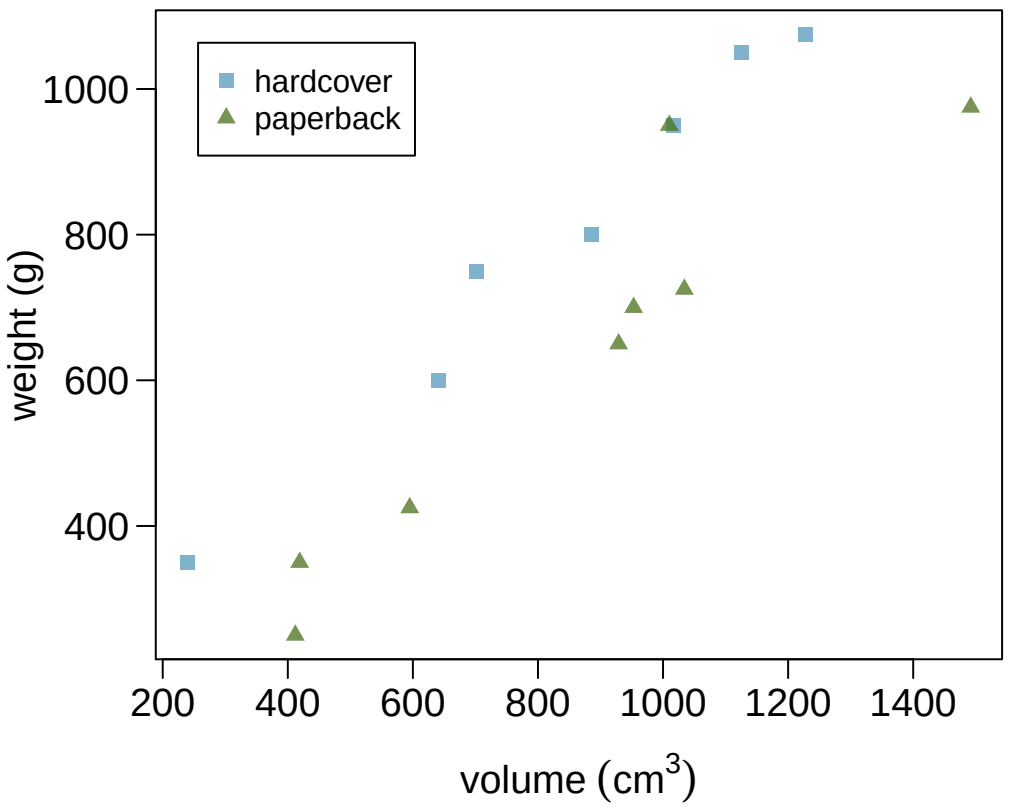
somewhat abbreviated output...

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.67931   88.37758   1.218    0.245
volume        0.70864    0.09746   7.271 6.26e-06

Residual standard error: 123.9 on 13 degrees of freedom
Multiple R-squared: 0.8026,Adjusted R-squared: 0.7875
F-statistic: 52.87 on 1 and 13 DF,  p-value: 6.262e-06
```

Weights of hardcover and paperback books

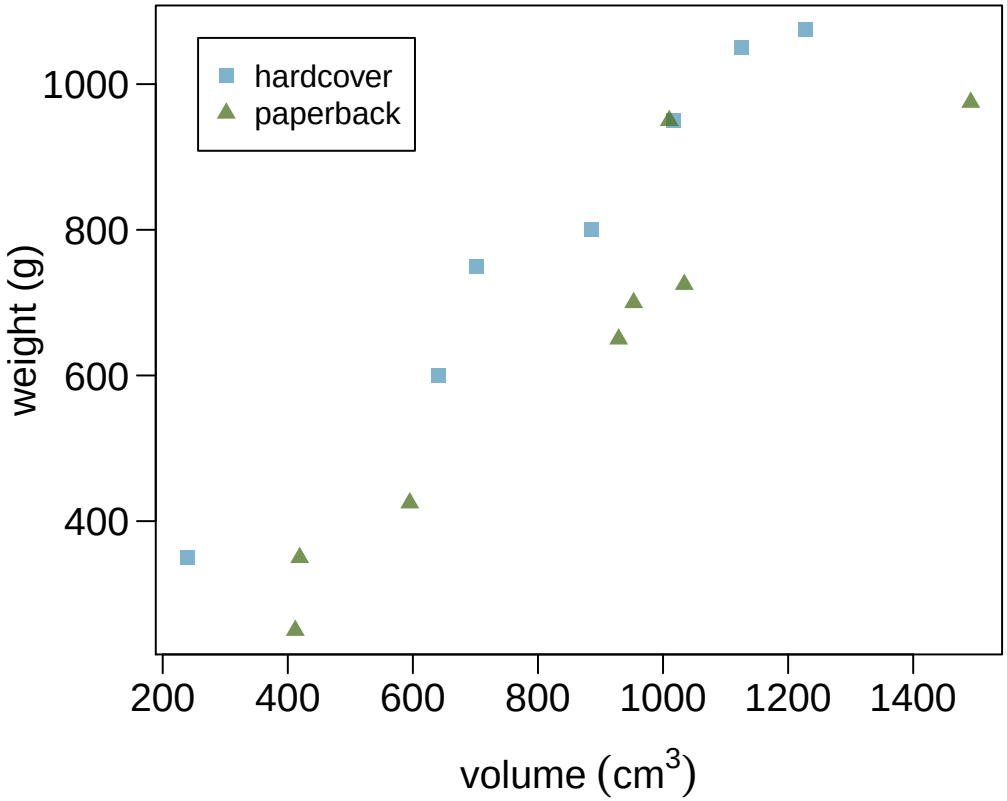
Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



Weights of hardcover and paperback books

Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?

Paperbacks generally weigh less than hardcover books after controlling for the book's volume.



Modeling weights of books using volume and cover type

```
book_mlr = lm(weight ~ volume + cover, data = allbacks)
summary(book_mlr)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  197.96284   59.19274   3.344 0.005841 **
volume         0.71795    0.06153  11.669  6.6e-08 ***
cover:pb    -184.04727   40.49420  -4.545 0.000672 ***


Residual standard error: 78.2 on 12 degrees of freedom
Multiple R-squared: 0.9275,Adjusted R-squared: 0.9154
F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```

Determining the reference level

Based on the regression output below, which level of *cover* is the reference level? Note that *pb*: paperback.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) paperback

(b) hardcover

Determining the reference level

Based on the regression output below, which level of *cover* is the reference level? Note that *pb*: paperback.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) paperback

(b) *hardcover*

Determining the reference level

Which of the below correctly describes the roles of variables in this regression model?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) response: weight, explanatory: volume, paperback cover

(b) response: weight, explanatory: volume, hardcover cover

(c) response: volume, explanatory: weight, cover type

(d) response: weight, explanatory: volume, cover type

Determining the reference level

Which of the below correctly describes the roles of variables in this regression model?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) response: weight, explanatory: volume, paperback cover

(b) response: weight, explanatory: volume, hardcover cover

(c) response: volume, explanatory: weight, cover type

(d) *response: weight, explanatory: volume, cover type*

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover : pb$$

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover : pb$$

1. For *hardcover* books: plug in *0* for *cover*

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 0$$

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover : pb$$

1. For *hardcover* books: plug in *0* for *cover*

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume \end{aligned}$$

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover:pb$$

1. For *hardcover* books: plug in *0* for *cover*

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume \end{aligned}$$

2. For *paperback* books: plug in *1* for *cover*

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 1$$

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover:pb$$

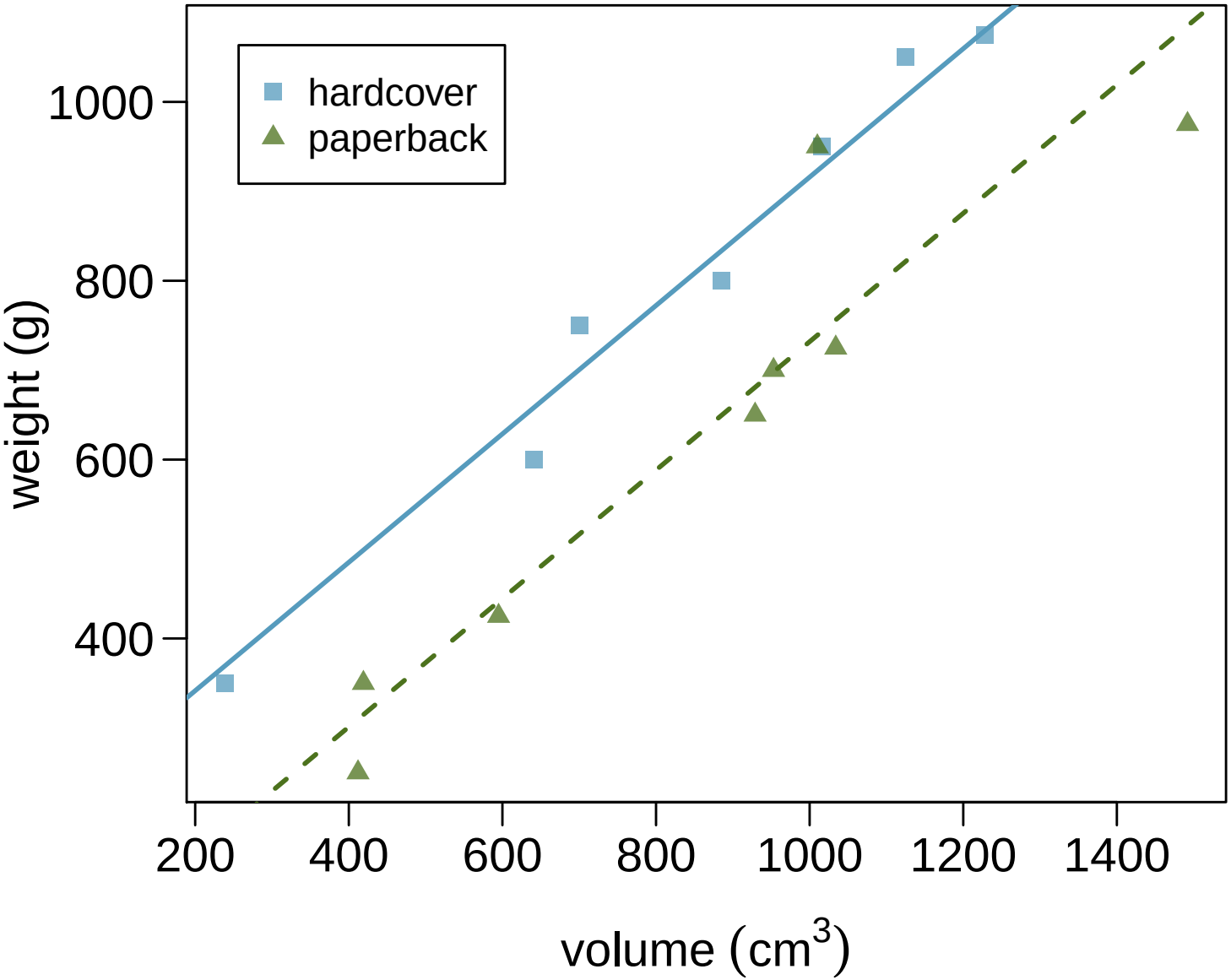
1. For *hardcover* books: plug in *0* for *cover*

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume \end{aligned}$$

2. For *paperback* books: plug in *1* for *cover*

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 1 \\ &= 13.91 + 0.72\ volume \end{aligned}$$

Visualising the linear model



Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.

Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.

Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.
- *Intercept:* Hardcover books with no volume are expected on average to weigh 198 grams.

Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.
- *Intercept:* Hardcover books with no volume are expected on average to weigh 198 grams.
  - Obviously, the intercept does not make sense in context. It only serves to adjust the height of the line.

Prediction

Which of the following is the correct calculation for the predicted weight of a paperback book that is 600 cm^3?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

(a) 197.96 + 0.72 * 600 - 184.05 * 1

(b) 184.05 + 0.72 * 600 - 197.96 * 1

(c) 197.96 + 0.72 * 600 - 184.05 * 0

(d) 197.96 + 0.72 * 1 - 184.05 * 600

Prediction

Which of the following is the correct calculation for the predicted weight of a paperback book that is 600 cm^3?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

(a) *197.96 + 0.72 * 600 - 184.05 * 1 = 445.91 grams*

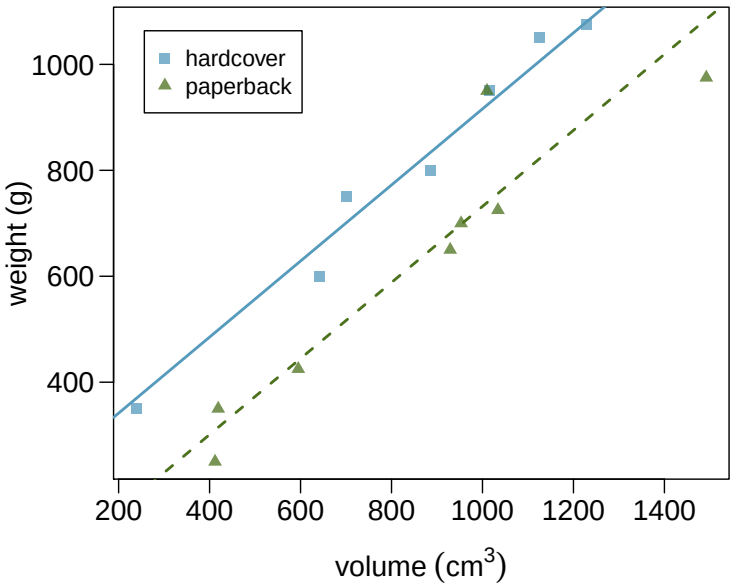
(b) 184.05 + 0.72 * 600 - 197.96 * 1

(c) 197.96 + 0.72 * 600 - 184.05 * 0

(d) 197.96 + 0.72 * 1 - 184.05 * 600

A note on “interaction" variables

$$\widehat{\text{weight}} = 197.96 + 0.72\text{volume} - 184.05\text{cover} : \text{pb}$$



This model assumes that hardcover and paperback books have the same slope for the relationship between their volume and weight. If this isn’t reasonable, then we would include an “interaction" variable in the model (beyond the scope of this course).

Another example: Modeling kid's test scores

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮					
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮					
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (2007) Cambridge University Press.

Interpreting the slope

What is the correct interpretation of the slope for mom's IQ?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

All else held constant, kids with mothers whose IQs are one point higher tend to score on average 0.56 points higher.

Interpreting the slope

What is the correct interpretation of the intercept?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

Interpreting the slope

What is the correct interpretation of the intercept?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

Kids whose moms haven't gone to HS, did not work during the first three years of the kid's life, have an IQ of 0 and are 0 yrs old are expected on average to score 19.59. Obviously, the intercept **does not make any sense** *in context.*

Interpreting the slope

What is the correct interpretation of the slope for *mom_work*?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

All else being equal, kids whose moms worked during the first three year's of the kid's life

(a) are estimated to score 2.54 points lower

(b) are estimated to score 2.54 points higher

than those whose moms did not work.

Interpreting the slope

What is the correct interpretation of the slope for *mom_work*?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

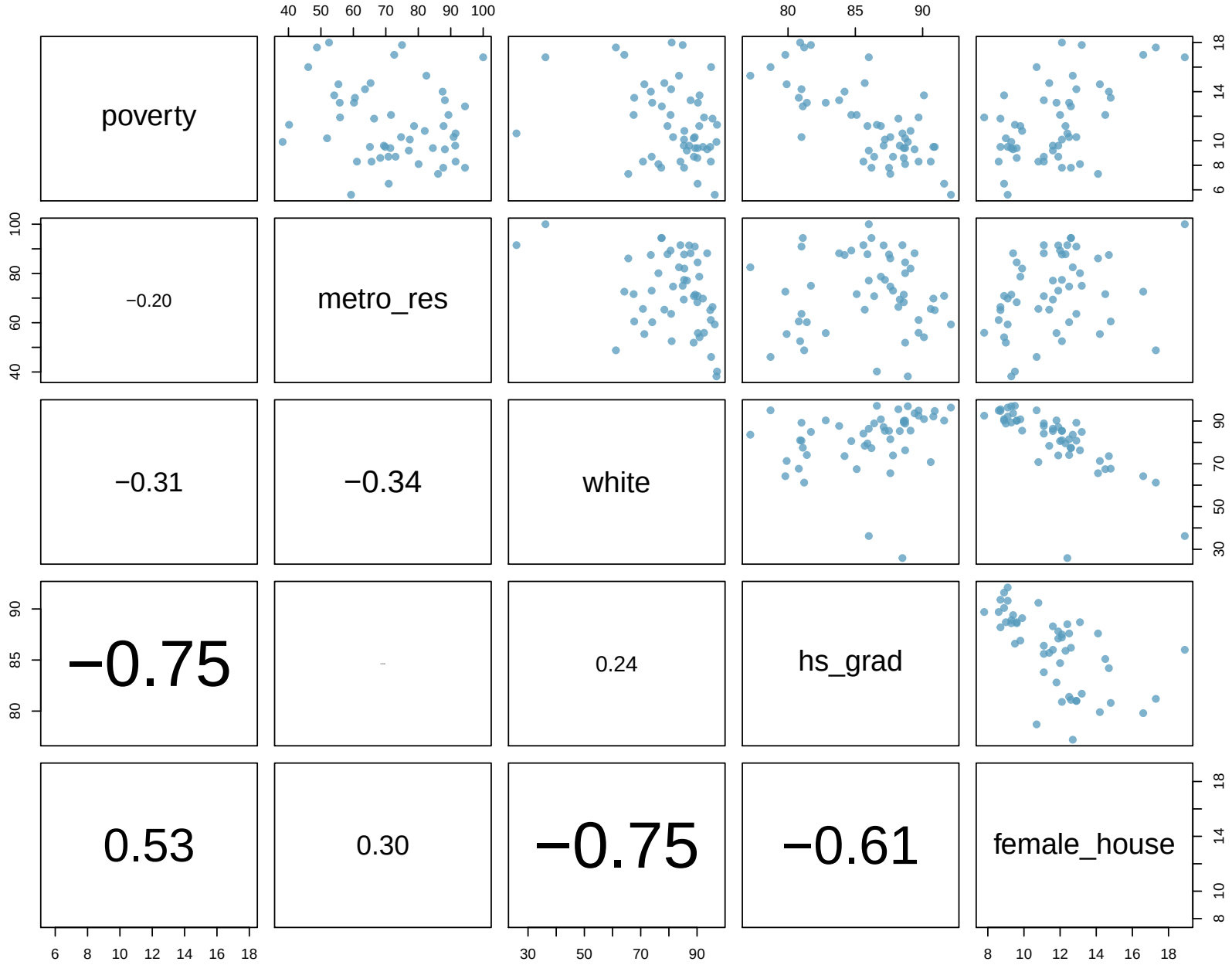
All else being equal, kids whose moms worked during the first three year's of the kid's life

(a) are estimated to score 2.54 points lower

(b) *are estimated to score 2.54 points higher*

than those whose moms did not work.

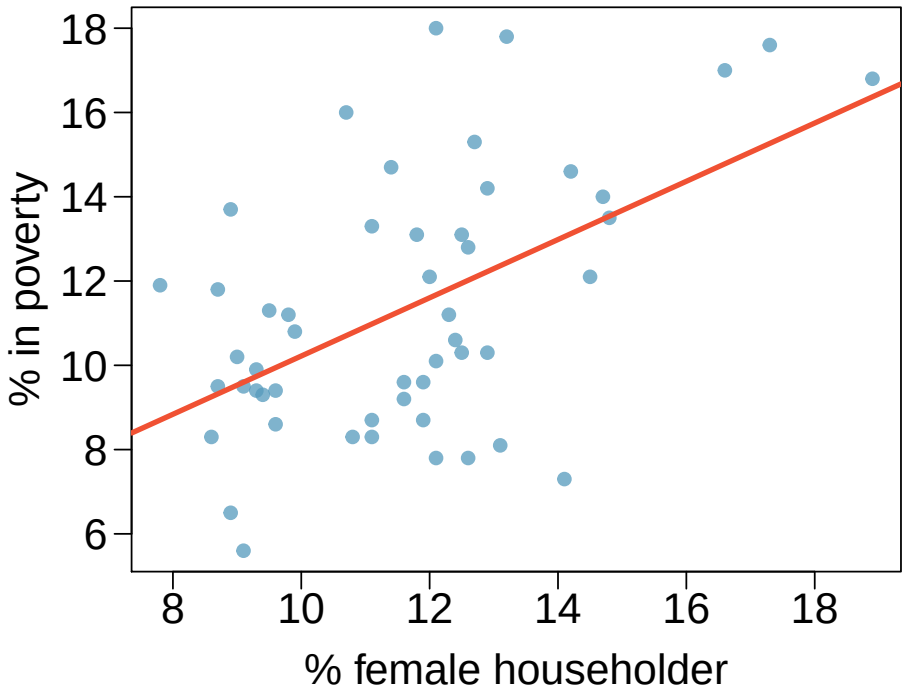
Revisit: Modeling poverty



Predicting poverty using % female householder

```
> pov_slr = lm(poverty ~ female_house, data = poverty)
> summary(pov_slr)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69	0.16	4.32	0.00



$$R = 0.53$$

$$R^2 = 0.53^2 = 0.28$$

Another look at R^2

R^2 can be calculated in three ways:

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$
3. based on definition:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$
3. based on definition:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

Using *ANOVA* we can calculate the explained variability and total variability in y.

Sum of squares

```
> anova(pov_slr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

Sum of squares

```
> anova(pov_slr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

$$\text{SS of } y\text{: } SS_{Total} = \sum (y - \bar{y})^2 = 480.25 \rightarrow \textit{total variability}$$

Sum of squares

```
> anova(pov_slr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

$$\text{SS of } y\text{: } SS_{Total} = \sum (y - \bar{y})^2 = 480.25 \rightarrow \textit{total variability}$$

$$\text{SS of residuals: } SS_{Error} = \sum e_i^2 = 347.68 \rightarrow \textit{unexplained variability}$$

Sum of squares

```
> anova(pov_slr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

$$\begin{aligned}
\text{SS of } y\text{: } SS_{Total} &= \sum (y - \bar{y})^2 = 480.25 \rightarrow \textit{total variability} \\
\text{SS of residuals: } SS_{Error} &= \sum e_i^2 = 347.68 \rightarrow \textit{unexplained variability} \\
\text{SS of } x\text{: } SS_{Model} &= SS_{Total} - SS_{Error} \rightarrow \textit{explained variability} \\
&= 480.25 - 347.68 = 132.57
\end{aligned}$$

Sum of squares

```
> anova(pov_slr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

$$\begin{aligned}
\text{SS of } y\text{: } SS_{Total} &= \sum (y - \bar{y})^2 = 480.25 \rightarrow \textit{total variability} \\
\text{SS of residuals: } SS_{Error} &= \sum e_i^2 = 347.68 \rightarrow \textit{unexplained variability} \\
\text{SS of } x\text{: } SS_{Model} &= SS_{Total} - SS_{Error} \rightarrow \textit{explained variability} \\
&= 480.25 - 347.68 = 132.57
\end{aligned}$$

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57}{480.25} = 0.28 \checkmark$$

Why bother?

Why bother with another approach for calculating R^2 when we had a perfectly good way to calculate it as the correlation coefficient squared?

Why bother?

Why bother with another approach for calculating R^2 when we had a perfectly good way to calculate it as the correlation coefficient squared?

- *For single-predictor linear regression, having three ways to calculate the same value may seem like overkill.*
- *However, in multiple linear regression, we can't calculate R^2 as the square of the correlation between x and y because we have multiple xs.*
- *And next we'll learn another measure of explained variability, adjusted R^2, that requires the use of the third approach, ratio of explained and unexplained variability.*

Predicting poverty using % female hh + % white

Linear model:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

Predicting poverty using % female hh + % white

Linear model:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57 + 8.21}{480.25} = 0.29$$

Adjusted R^2

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

where n is the number of cases and p is the number of predictors (explanatory variables) in the model.

- Because p is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors included in the model.
- Therefore, we choose models with higher R^2_{adj} over others.

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right)
\end{aligned}
$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right)
\end{aligned}
$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right) \\
&= 1 - 0.74
\end{aligned}
$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right) \\
&= 1 - 0.74 \\
&= 0.26
\end{aligned}
$$

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

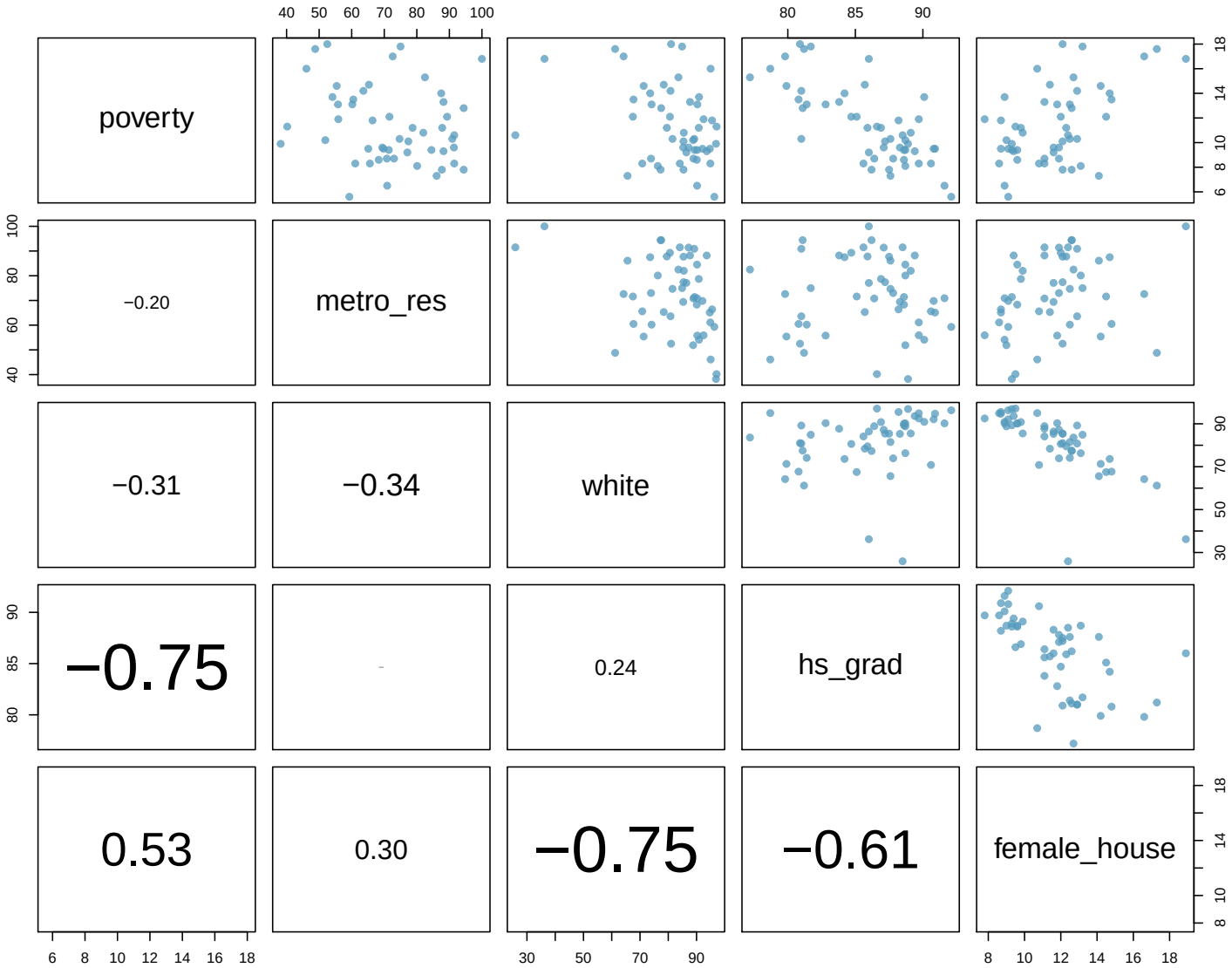
- When any variable is added to the model R^2 increases.

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

- When any variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

We saw that adding the variable *white* to the model did not increase adjusted R^2, i.e., did not add valuable information to the model. Why?



Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.

Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.

- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.

Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.

- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.
- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to prevent correlation among predictors.

Variable Selection

In practice, we often have a long list of potential explanatory variables and need to make a decision on which variables to include.

As an example, the Princeton economist Orley Ashenfelter built a model to predict the price of wine, along the following lines:

$$\begin{aligned} \textbf{price}_i &= \beta_0 + \beta_1(\textbf{avg. rainfall})_i + \beta_2(\textbf{avg. temp.})_i + \\ &\quad \beta_3(\textbf{calcium in soil})_i + \beta_4(\textbf{soil pH})_i + \epsilon_i \end{aligned}$$

This general kind of model is often used by wine speculators.

In building such a model, Ashenfelter considered many possible explanatory variables. He wanted to include only those that were relevant. If the model includes irrelevant explanatory variables, then it tends to give poor predictions.

To determine which variables to include and which to remove from his model, Ashenfelter did hypothesis tests to decide whether each estimated coefficient was significantly different from zero.

Variable Selection

To make this test, the null and alternative hypotheses are:

$$\mathbf{H}_0 : \beta_i = 0 \textbf{ vs. } \mathbf{H}_A : \beta_i \neq 0.$$

The test statistic takes the form we are used to:

$$ts = \frac{pe - 0}{se} = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\beta_i}}$$

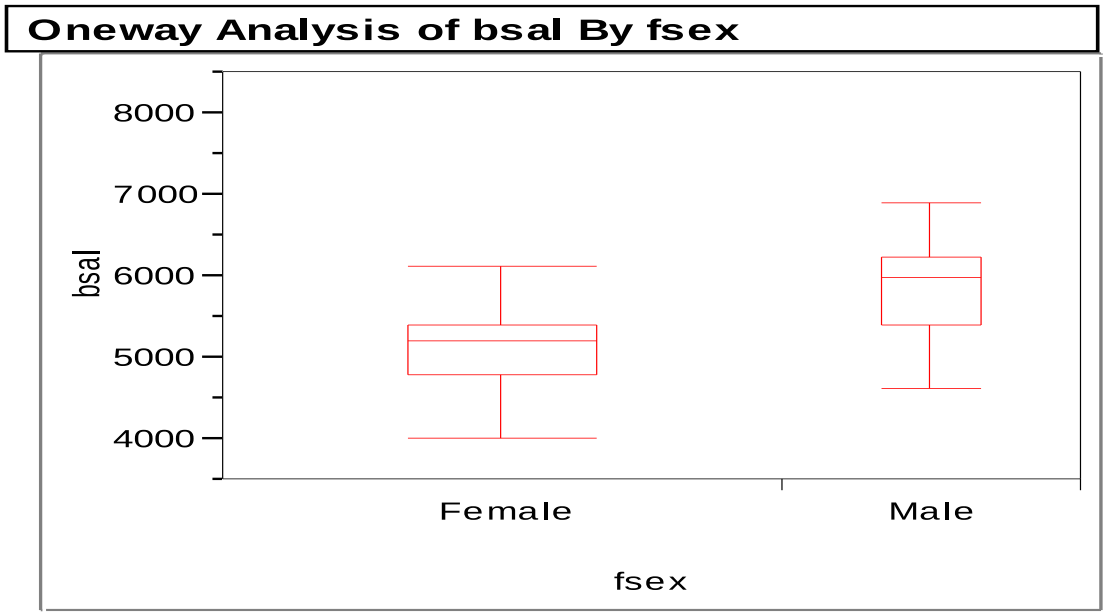
where $\hat{\sigma}_{\beta_i}$ is the standard error of the estimate $\hat{\beta}_i$. It is a bit complicated (remember that we found the variance of $\hat{\beta}_1$ in the previous lecture), but can be found from the all standard statistics packages.

This ts is compared to a t-distribution with $n - p - 1$ degrees of freedom (we lose information equivalent to one observation for each parameter we estimate, and we had to estimate $\beta_0, \ldots, \beta_p$). If $n - p - 1 > 30$, we can use the z-table.

Examples

Why multiple regression? This follows intuitively from our discussion on confounders and the following example illustrates the point.

Example 1: In 1979, Harris Trust and Savings Bank was accused of gender discrimination in starting salaries. In particular, one main question was whether men in entry-level clerical jobs got higher salaries than women with similar credentials. Exploratory box plots showed that the claim might be true.



Examples

Harris Trust and Savings denied that they discriminated. They claimed that their starting salaries were based on many other factors, such as seniority, education, age and experience (possible confounders).

To assess that claim, the plaintiffs' lawyers used multiple regression:

$$\begin{aligned}\textbf{salary}_i &= \beta_0 + \beta_1(\textbf{sex})_i + \beta_2(\textbf{seniority})_i + \beta_3(\textbf{age})_i + \beta_4(\textbf{educ})_i + \\ &\quad \beta_5(\textbf{exper})_i + \epsilon_i\end{aligned}$$

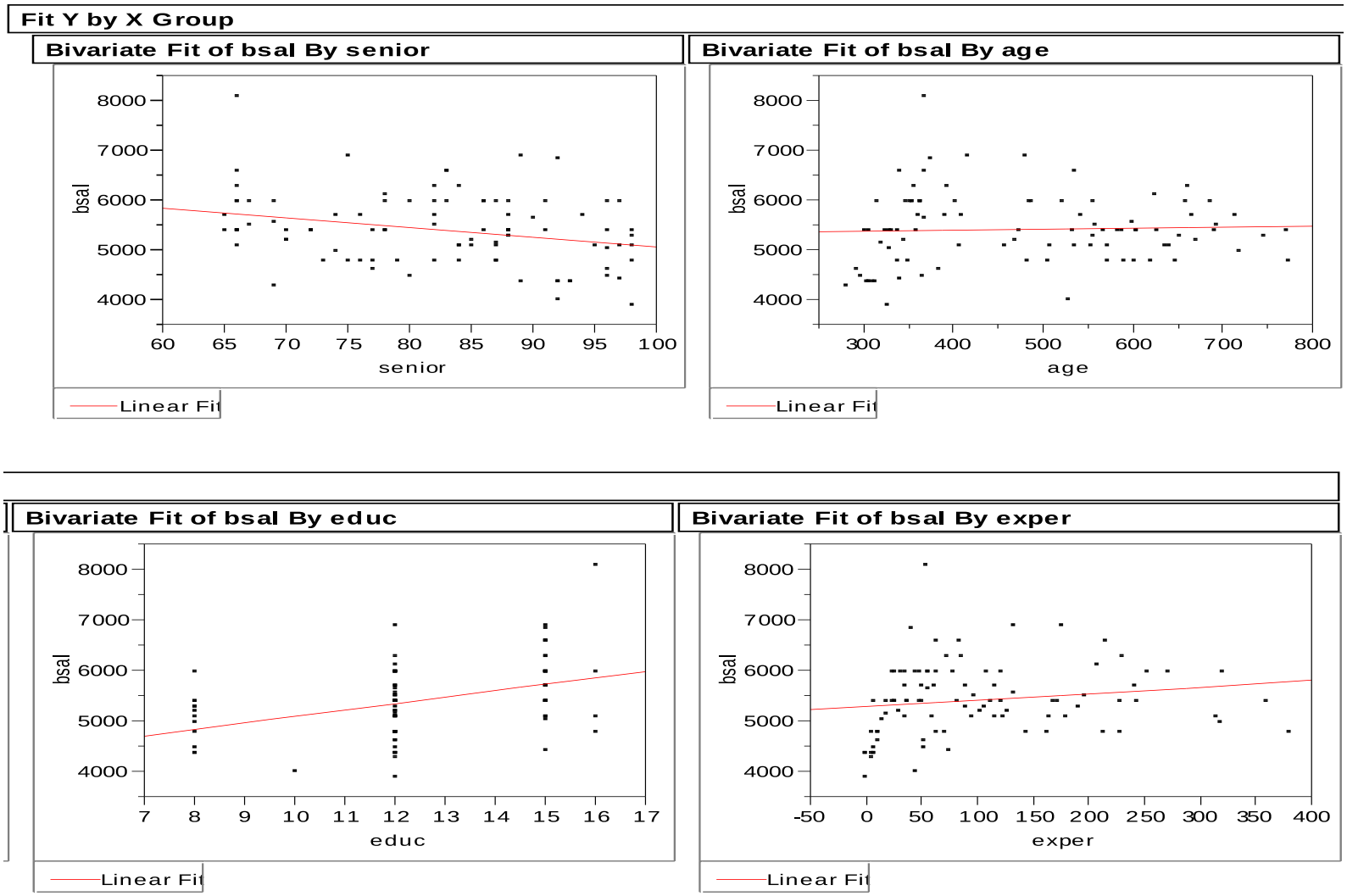
Sex was recorded as 1 if the person was female, 0 for males.

Age, seniority, and experience were measured in months. Education was measured in years (we are treating education as a numeric variable but that's not the only way to treat it; let's ignore that discussion here though).

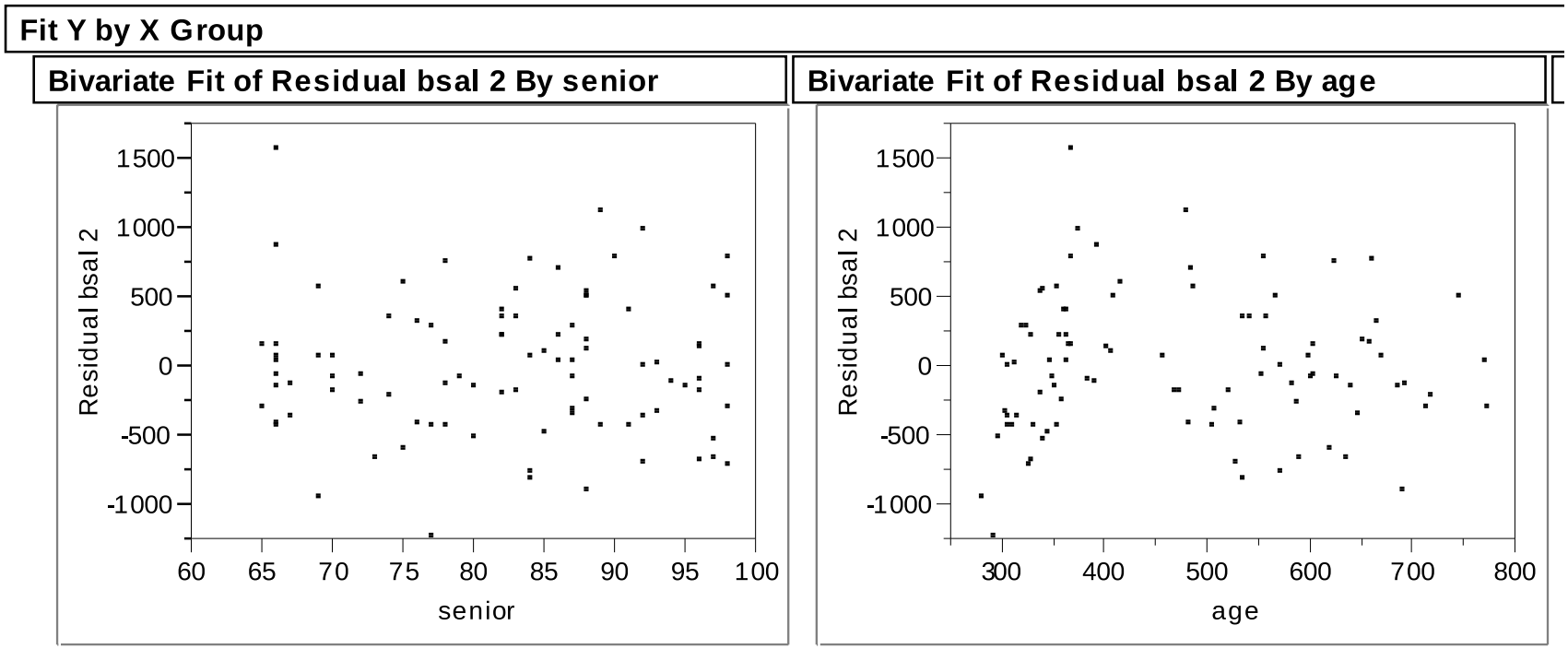
The legal question was whether the coefficient β_1 was significantly less than 0. If so, then the effect of gender was to lower the starting salary.

Examples

Let's look at the regression fit by each variable, holding the rest constant.

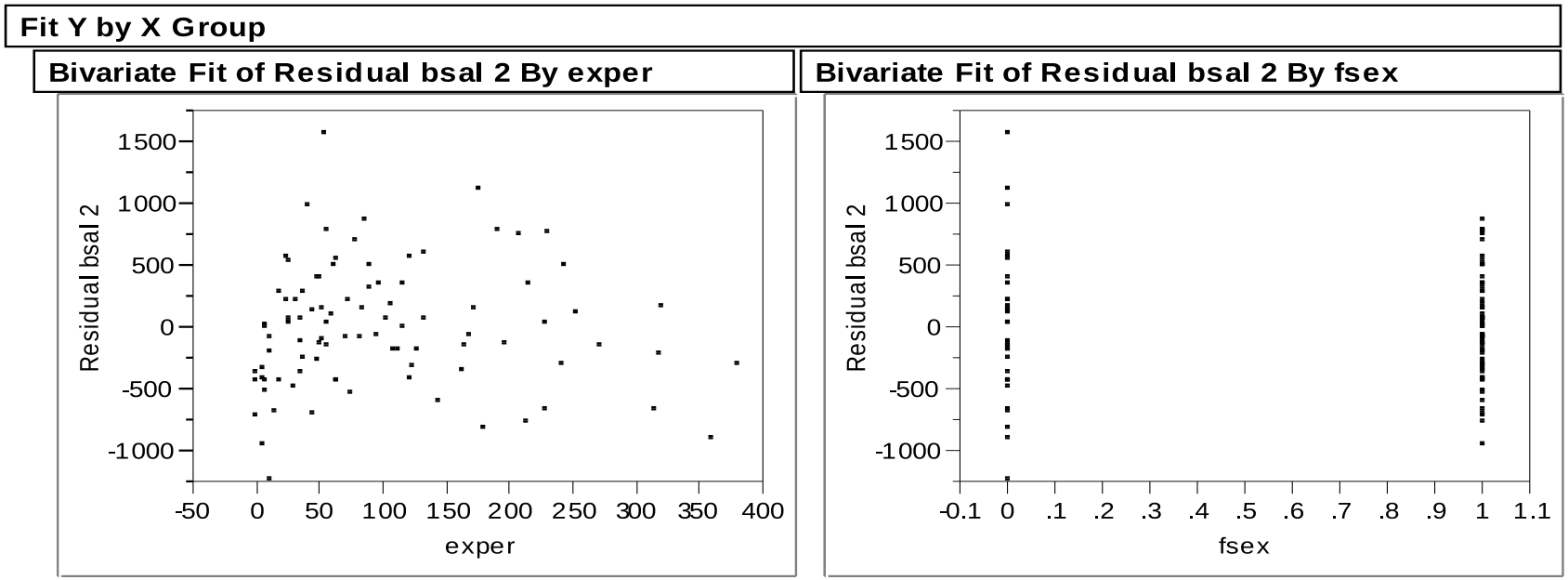


Examples



These are some of the residual plots, to examine our assumption of independence between the errors and each covariate. The seniority plot looks pretty good, there is something at little odd for age at around 400 months (age 33).

Examples



These are more residual plots. Experience may show some patterning. Gender shows that there is more variance for men than for women. One residual may be the boss's son?

Residual plots are really useful in examining our assumption of independence. They might help us figure out that we are still missing a function of one of the x's if we observe any pattern.

Examples

Back to the question we are interested in answering:

Using the 93 available cases of entry-level clerical workers, a statistical package found that the estimated model is

$$\begin{aligned}\textbf{salary}_i &= 6277.9 - 767.9(\textbf{sex})_i - 22.6(\textbf{seniority})_i + 0.63(\textbf{age})_i + \\ &\quad 92.3(\textbf{educ})_i + 50(\textbf{exper})_i + \epsilon_i\end{aligned}$$

The output showed that the standard error for the estimate of the coefficient on sex (i.e., the $\hat{\sigma}_{\beta_1}$) was 128.9.

We observe that the coefficient on sex is negative, which suggests that there may be discrimination against women. But we still need a significance test to be sure. We cannot interpret the size of the effect without one. Without a small p-value (below $\alpha = 0.05$ for example), Harris Trust and Savings might argue in court that this result is only due to random chance.

Examples

Because we care about a one-sided alternative hypothesis, the null and alternative hypotheses are:

$$\mathbf{H}_0 : b_1 \geq 0 \textbf{ vs. } \mathbf{H}_A : b_1 < 0.$$

The test statistic is

$$ts = \frac{\hat{b}_1 - 0}{se} = \frac{-767.9 - 0}{128.9} = -5.95.$$

This is compared to a t-distribution with $n - p - 1 = 93 - 5 - 1 = 87$ degrees of freedom. Since this is off our t-table scale, we use a z-table. The result is highly significant. Reject the null; there is evidence of discrimination.

Recap

Today we learned about extending the regression idea to multiple explanatory variables, and the adjusted R^2.

Suggested reading:

- D.S. Sec. 11.5
- OpenIntro3: Sec. 8.1