Elementary Regression 1

Paul E. Johnson[1] [2]

[1]Department of Political Science

[2]Psychology

2020

Elementary OLS

Outline

Data Set: Columns of Same Length

row number	respondent id	*income*	*educ*	gender
1	243223	4352.5	6	M
2	151512	6525.1	21	F
3	515131	4345.5	13	M
4	166122	3421.4	12	F
⋮				

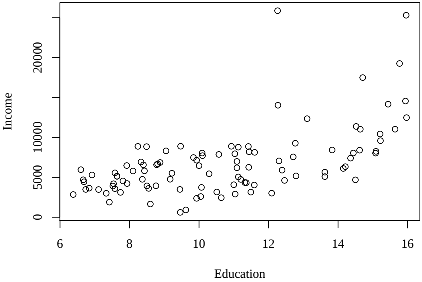
- Variables are “columns” in a data frame
- Rows are called “observations” or “cases” or “respondents” or “subjects”
- Talk about row “i” if you mean to say something that applies for each row

Design Matrix

- Regression is, inherently, a procedure for estimating effects of numeric predictors
- The data frame (in R, the "model frame") has to be converted from data as we see it into a thing that has only numeric columns.
- Categorical predictors are converted into "indicator" variables (dummy variables, usually coded {0,1} or {-1,1}

row number	respondent id	*income*	*educ*	gender
1	243223	4352.5	6	1
2	151512	6525.1	21	0

Scatterplot: One Input, One Output



.

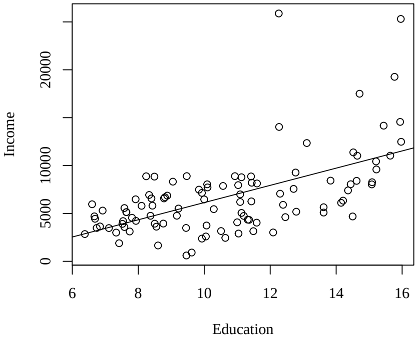
.

The “Prestige” dataset in the R package “car” by John Fox

Dependent, Independent

- DV: Dependent Variable: The thing we are predicting
  - The "output" variable, generally we call it y_i
    - In this case "$income_i$".
  - Synonyms: "endogenous variable" "outcome variable"
- IV: Independent Variable
  - The "input" variable, generally call it x_i,
    - In this case "$education_i$" .
  - Synonyms: "exogenous variable" "predictor" "covariate"
- Regression allows several input variables, but for now we consider only one.

Line of Best Fit



- This is the Straight Line that "best fits" the data
- Best fit = minimizes a criterion, here the "sum of squared errors"
- "Predicted value" synonym for "fitted value" or "conditional expected value"
  - For any value of *education*, we predict an outcome on the line
- Later, we will use diagnostics to test suitability of this model

Typical Computer Printout Summarizing a Fitted Regression

```
Call:
lm(formula = income ~ education, data = Prestige)

Residuals:
    Min       1Q   Median       3Q      Max
-5493.2  -2433.8    -41.9   1491.5  17713.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2853.6     1407.0  -2.028   0.0452 *
education       898.8      127.0   7.075 2.08e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3483 on 100 degrees of freedom
Multiple R^2:  0.3336, Adjusted R^2:  0.3269
F-statistic: 50.06 on 1 and 100 DF,  p-value: 2.079e-10
```

Make a Professionally Acceptable Regression Table

	M1	
	Estimate	(S.E.)
(Intercept)	-2853.586*	(1407.039)
education	898.813***	(127.035)
N	102	
RMSE	3483.378	
R^2	0.334	

$*p \leq 0.05{**}\ p \leq 0.01{***}p \leq 0.001$

- When we are finished, you will understand all of these details.

In R, after "lm", run follow-up functions

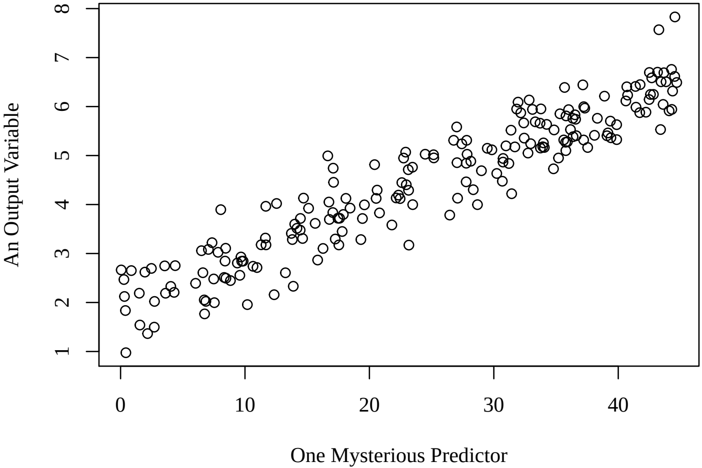
There are many (at least 30) "methods" that can be used to explore that fitted model.

```
incedmod1 <- lm(income~education,
     data=Prestige)
summary(incedmod1)
anova(incedmod1, test="F")
vcov(incedmod1)
confint(incedmod1)
plot(incedmod1)
termplot(incedmod1, se=T, partial=T)
```

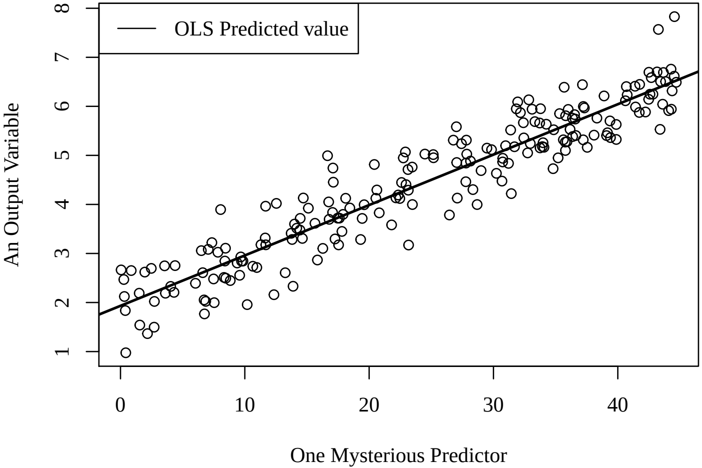
- lm: creates the regression model "incedmod1"
- summary: main regression table
- anova: asks for sum of squares information
- vcov: asks for the variance/covariance matrix of $\hat{\beta}$'s
- confint: confidence intervals for intercept and slope
- plot: creates diagnostic displays
- termplot: plots the predictive line
- many methods in the "car" package
- rockchalk plotting and diagnostic routines

Outline

1. Can I Run Regression on This?

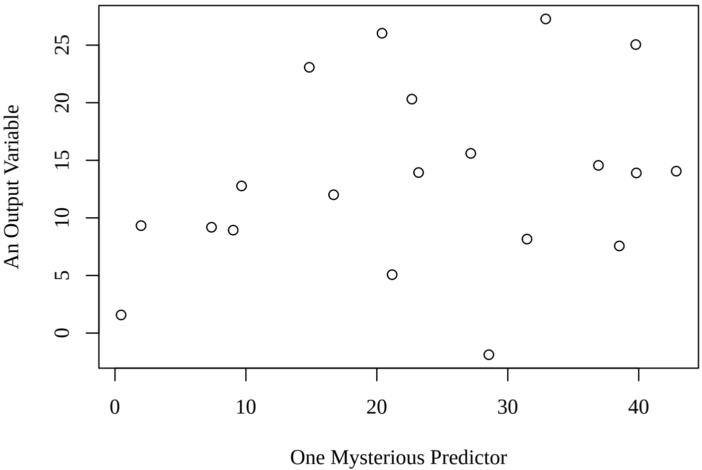


1. As we say in Francais, Oui!

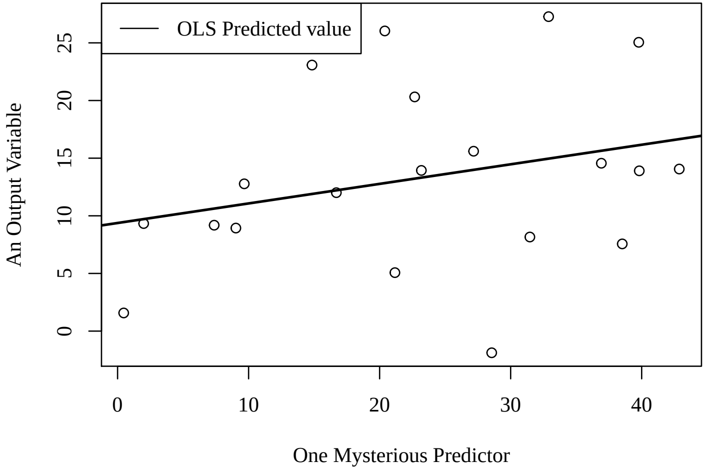


2 numeric variables, passes the “inter-occular trauma test”

2. Can I Run Regression on This?

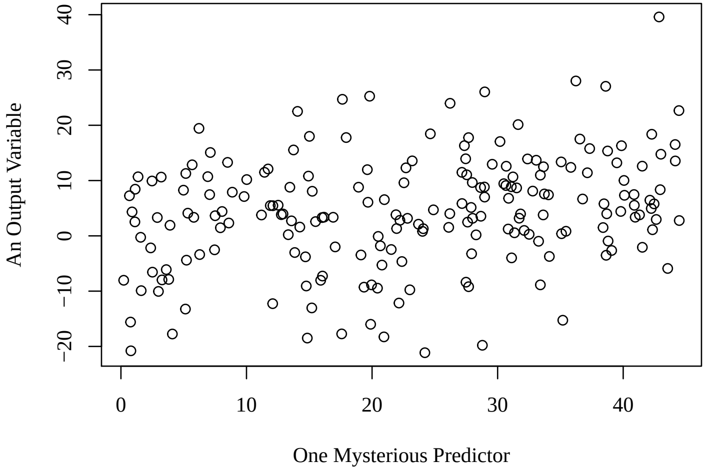


2. Sure, Why Not?

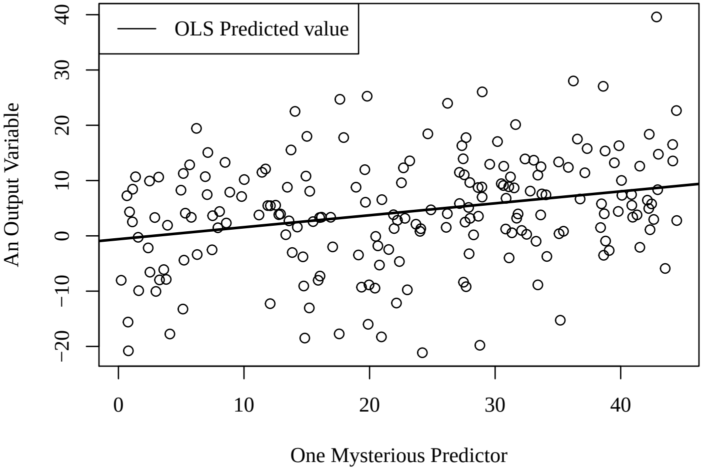


The "straight line" prediction is not wrong. But not precise, either.

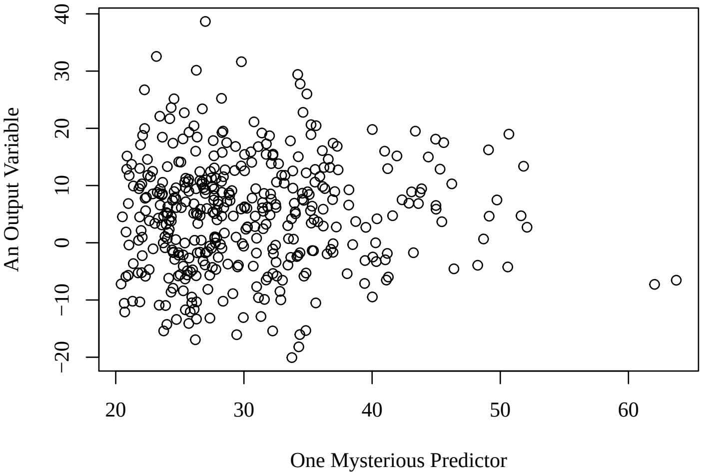
3. Can I Run Regression on This?



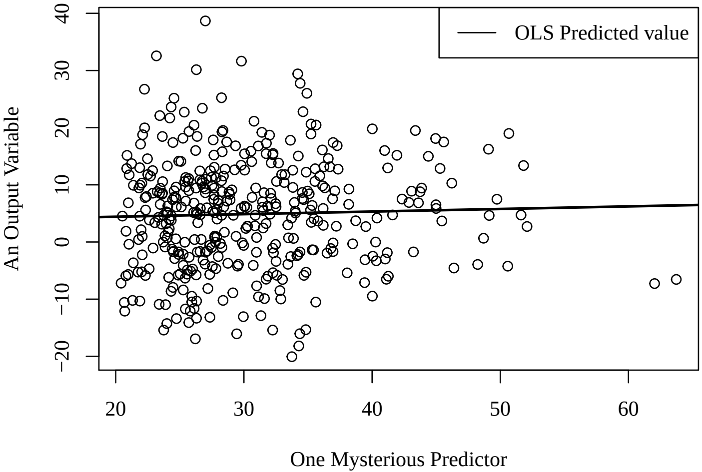
3. En Espanol, Si!



4. Can I Run Regression on This?

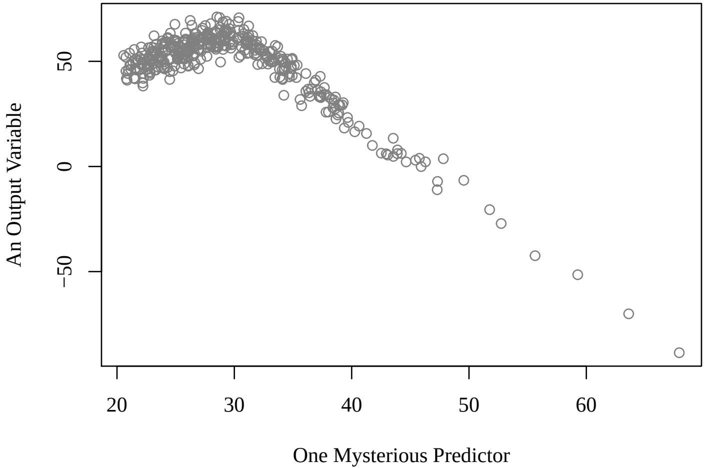


4. OK, I Don't Mind a Bit

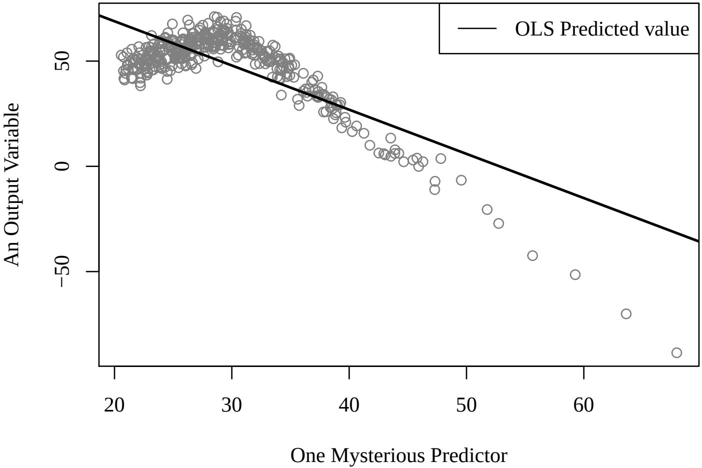


I don't know of any reason why you expect the predictor to be "evenly distributed" or "normal" or whatnot

5. Can I Run Regression on This?



5. No. Are You Joking?

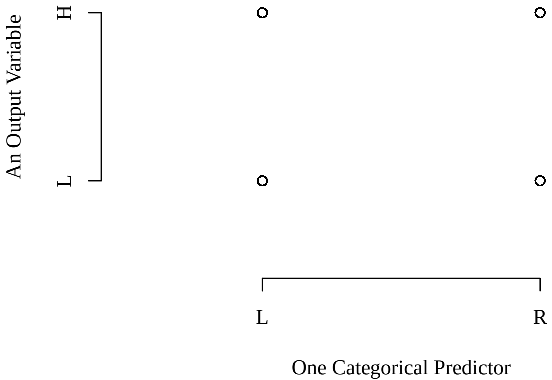


Straight line does not suit this data

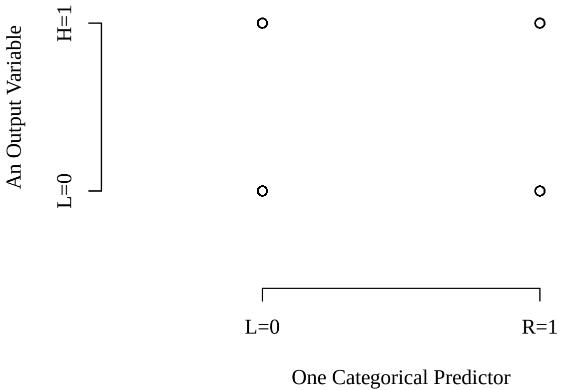
What's the point so far?

- We don't assume much about the predictor
- We do assume a LOT about the outcome variable
  - it is supposed to be scattered "equally likely" above and below the line

6. Can I Run Regression on This?

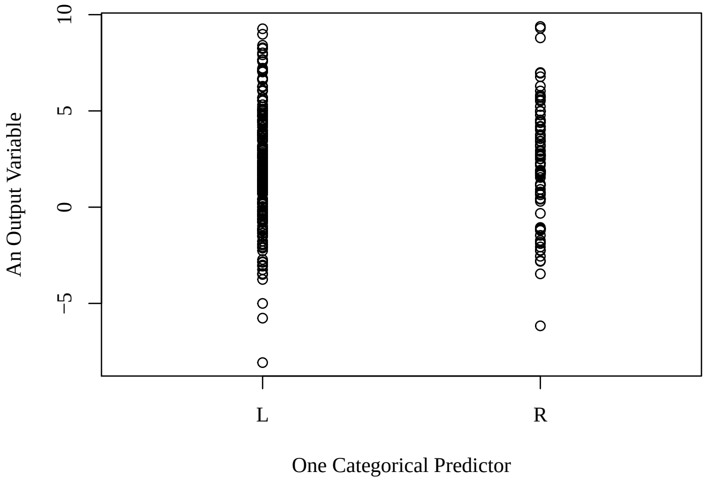


6. Maybe, But You'd Really Have to Stretch

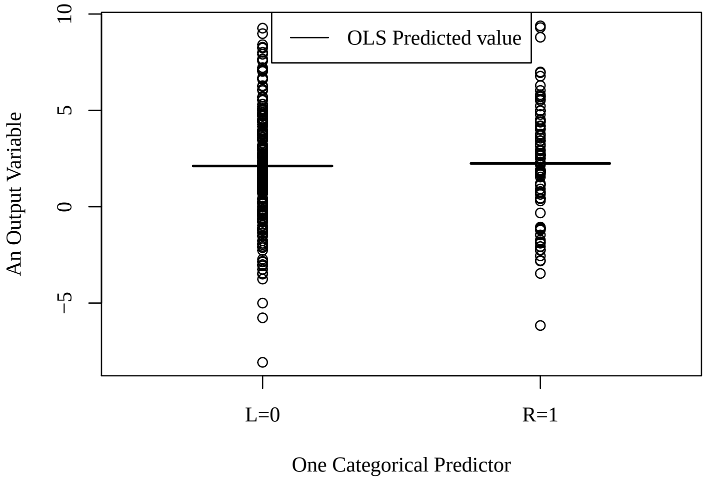


Its tough for me to see a "regression line" in there, but some people do.

7. Can I Run Regression on This?

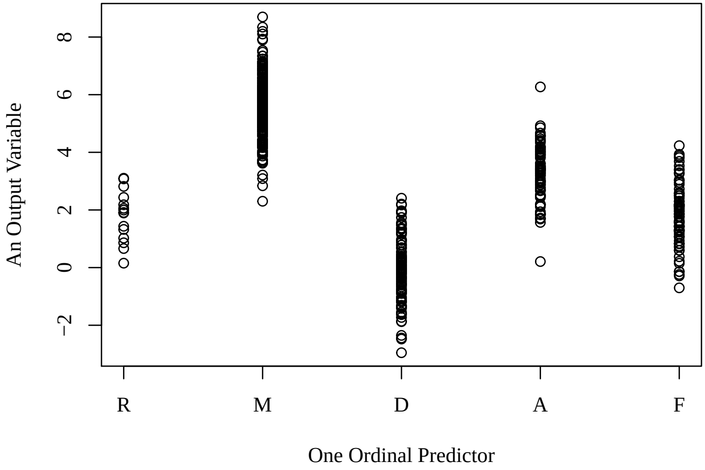


7. Probably, if you recode the predictor as {0,1}

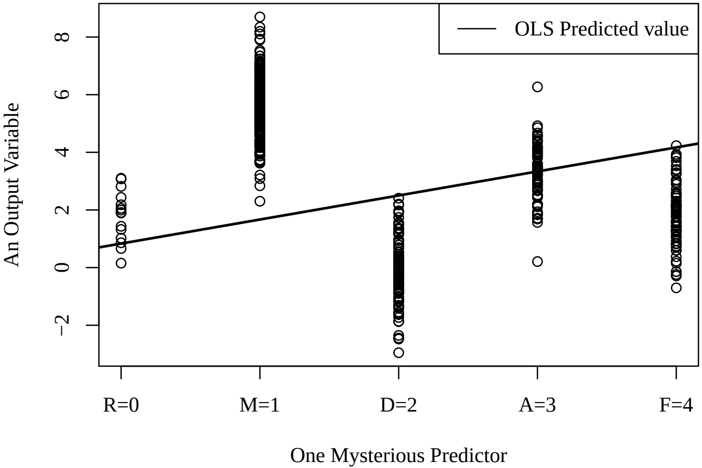


The appropriate graph has "steps", rather than a line. Predictions for discrete points.

8. Can I Run Regression on This?



8. As Yoda says, “Mistaken, This Appears”



Outline

Assumption 1: Linear Relationship

- For each "case" i, the following is true:

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{1}$$

- The parameters are β_0, β_1, and σ_e
  - β_0 is the "constant" or "y intercept".
  - β_1 is the slope of the line.
  - σ_e is the standard deviation of a "random effect," e_i, that is uniquely drawn for each observation.
- The subscript i means x_i and y_i are individual specific. Note no i on β's or σ_e
- In the past, my notes used the letter b for coefficients, not β, mostly because b was easier to type in MS Word. Now I use LaTeX, I don't have that problem anymore. But I have not updated all of my notes about everything.

Random and Deterministic Parts

- The deterministic part is the "true line" $\beta_0 + \beta_1 x_i$
- The stochastic (random part) "throws" observed scores up and down

0_home_pauljohn_SVN_SVN-guides_stat_R

Separate Deterministic and Stochastic Parts

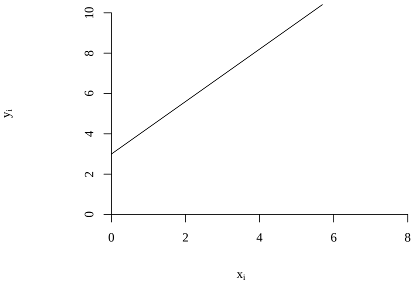
- Suppose $\beta_0 = 3$ and $\beta_1 = 1.3$.
- The "true relationship":

$$y_i = 3 + 1.3 \cdot x_i + e_i$$

- The deterministic part:

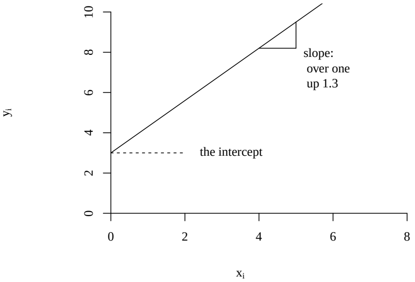
$$3 + 1.3 \cdot x_i$$

- The stochastic part is e_i.



Refresher: Linear Equation

- $3 + 1.3 \cdot x_i$
- The slope: 1.3 is the "rise over run"
  - For each 1 unit increase in x_i, the outcome increases by 1.3.
- The intercept: 3
  - When $x_i = 0$, the outcome will be 3.

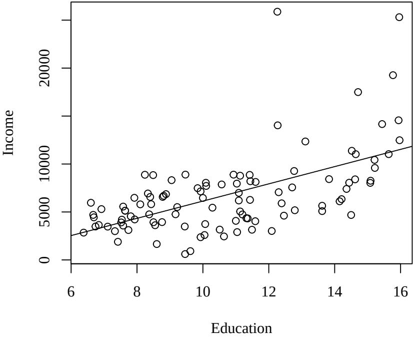


The Fitted Line in the Income Equation

- Note the difference between the theory and the estimate
- Theory:
  $income_i = \beta_0 + \beta_1 education_i + e_i$
- Estimated line:

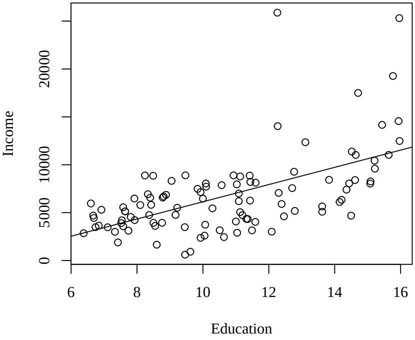
$$\widehat{income}_i = -2853.585 + 898.813 \cdot education_i \quad (2)$$

- There is no "error term" in the equation for the predicted line. That's because we assume $E[e_i] = 0$.



The Fitted Line in the Income Equation

- A 1 unit increase in $education_i$ "is associated with"(causes?) a 898.8 increase in $income_i$
- The subscript i is important. It helps us remember the assumption that the same relationship applies for all cases, $i \in \{1, \ldots, N\}$
- The regression model also summarizes the "scatter" above and below, which is our next topic.



Assumption 2: A "Well Behaved Error Term"

- We don't have to say e_i is $Normal(0, \sigma_e^2)$. But we could. Some people do.
- Well behaved means "symmetric" and "homogeneous", which is not as strong as assuming Normal
  - Assumption 2A: e_i is "on average" 0: $E[e_i] = 0$
  - Assumption 2B: all observations are drawn from the same distribution with a constant variance, σ_e^2 (a.k.a "homoskedasticity")

$$Var[e_i] = E[e_i^2] = \sigma_e^2$$

- Violations of these assumptions lead to re-specification and advanced model-fitting techniques (nonlinear models, weighted least squares, random effects models)

Assumption 2A: $E[e_i] = 0$

- The error term has an average value of 0:

$$E(e_i) = 0 \tag{3}$$

- Thus $E[y_i|x_i] = E[\beta_0 + \beta_1 x_i + e_i] = \beta_0 + \beta_1 x_i + 0$
- You can guess where this leads, right?
  - If we had reasonable estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, the predicted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is a reasonable estimate of the expected value, given x_i.
  - In other words, it is not ridiculous to use predicted (or fitted) value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ as an estimated value for y_i

Assumption 2B: Homoskedasticity

- The error term's variance is constant, i.e, the same for all cases i

$$Variance[e_i] = \sigma_e^2 \tag{4}$$

- I.e., σ_e^2 **is the same for all cases**. It is not subscripted by i.
  - Every case's "random effect" comes from a distribution with the same amount of uncertainty in it.
- This assumption is vital in our understanding of uncertainty, or variance, in the estimates.

Sidenote. Explain $E[e_i^2] = \sigma_e^2$

- The Variance of the error term equals the expected value of e_i^2.
- Many stats book will define "homogeneous variance" as:

$$E[e_i^2] = \sigma_e^2$$

rather than the more obvious

$$Var[e_i] = \sigma_e^2 \tag{5}$$

- While disconcerting, we can show these are the SAME definitions. Start with the definition of variance

$$V(e_i) = \sigma_e^2 = E[(e_i - E[e_i])^2]$$

- Recall $E(e_i) = 0$, so

$$V[e_i] = E[(e_i - 0)^2] = E[e_i^2]$$

In Maximum Likelihood Analysis, A Stronger Assumption Would be Required

- In ML (including the generalized linear model), we would assume a specific distribution for e_i, which amounts to saying that we can state the distribution of y_i given x_i and the β's.
- We would usually say y_i, depends on "linear predictor" $(\beta_0 + \beta_1 x_i)$.
- For example, given x_i, y_i is Normal, i.e., drawn from $N(\beta_0 + \beta_1 x_i,\ \sigma_e^2)$
- Until the end of this class, we don't need to make that assumption, but you can if you like it!
- When you get to GLM, you can assume that y_i is Poisson, Gamma, or whatever you like.

Roadmap of Ahead

1. calculate estimates of β_0 and β_1 (which we will call $\hat{\beta}_0$ and $\hat{\beta}_1$)
2. evaluate our uncertainty about the $\hat{\beta}$'s by calculating standard errors of the $\hat{\beta}$.
3. estimate the variance of e_i, $\widehat{\sigma_e^2}$
4. conduct some "diagnostics" to find out if we might fit a better model.

Outline

Treat $\hat{\beta}_0$ and $\hat{\beta}_1$ as unknowns.

- This week, we only use a "straight line" predicted value formula.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i \tag{6}$$

- The observed variables x_i and y_i are now treated as "known values",
- The parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ become variables that we adjust to find the best prediction.

OLS: The Sum of Squares as a Criterion

- Predicted: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Residual: Difference between observed y_i and predicted $\hat{y}_i$.
- $S(\hat{\beta}_0, \hat{\beta}_1)$:Sum of Squared Residuals depends on $\hat{\beta}_0$, $\hat{\beta}_1$

$$\begin{aligned} S(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{N} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \\ &= \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned} \quad (7)$$

- OLS Criterion: minimize the sum of squared residuals by adjusting $\hat{\beta}_0$ and $\hat{\beta}_1$
- Notation alert: Often also called "sum of squared errors", but better to be clear: we never know "true errors", we only know "residuals". So I'm trying to remember to call it sum of squared residuals.

Estimation process is outlined in the Appendix

- The sum of squared residuals is an objective function that we minimize by adjusting $\hat{\beta}_0$ and $\hat{\beta}_1$
- Because the sum of squares is a "U" shaped function, we can visualize the solution.

1_home_pauljohn_SVN_SVN-guides_stat_Reg

The Solutions are the "OLS Estimators"

- We'd ordinarily use matrix algebra to solve this problem, but I don't want to go into matrices at this point.
- Thus I write out the solution in "scalar" format, using ordinary summations and such.

$$\hat{\beta}_1^{OLS} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{8}$$

$\bar{x}$ and $\bar{y}$ are sample means.

- Note
  - numerator terms: product of x deviations and y deviations about their means
  - denominator terms: x deviations squared.
- If you have "mean centered data", this simplifies to

$$\hat{\beta}_1^{OLS} = \frac{\sum x_i y_i}{\sum x_i^2} \tag{9}$$

The Solutions are the "OLS Estimators" ...

- And the intercept estimate: $\hat{\beta}_0^{OLS} = \bar{y} - \hat{\beta}_1^{OLS} \bar{x}$
- If you were paying attention when we studied Variance and Covariance, you notice the formula for $\hat{\beta}$ is $Cov(x, y)/Var(x)$. Interesting co-incidence, there.

Gauss Markov Theorem: OLS is B.L.U.E.

- $\hat{\beta}^{OLS}$ is an Unbiased estimator, it is "on average" correct: $E[\hat{\beta}^{OLS}] = \beta$
- $\hat{\beta}^{OLS}$ is Consistent, as $N \to \infty$, $\hat{\beta}^{OLS}. \to \beta$. (the probability that the gap $|\hat{\beta}^{OLS} - \beta|$ is bigger than any small number shrinks toward 0 as $N \to \infty$).
- $\hat{\beta}^{OLS}$ is Efficient: No linear unbiased estimating formulae has lower variance than $\hat{\beta}^{OLS}$.

Outline

Define residual, as opposed to "error"

- e_i is an "error term", it is unmeasured, unknown.
  - Its "true mean" (expected value) is assumed to be 0
  - Its "true variance" is σ_e^2, also unknown.
- $\hat{e}_i$ is the "residual", $y_i - \hat{y}_i$. It serves as an estimate of the error term.

MSE=Mean Square Error

- Predict $\hat{y}_i$ from the best fitting model
- The commonly-called MSE (Mean Squared Error) is the mean of squared **residuals**.

$$MSE = \frac{\sum(y_i - \widehat{y}_i)^2}{N-2} = \frac{\sum \widehat{e}_i^{\,2}}{N-2} \tag{10}$$

- MSE = unbiased estimator of σ_e^2 (because of $N-2$ in denominator). Unbiased means

$$E[MSE] = \sigma_e^2 \tag{11}$$

- Other notation for MSE: $\widehat{\sigma_e^2}, \widehat{Var[e_i]}$, s_e^2

RMSE=Root Mean Squared Error

- RMSE (root MSE) is the SAS name for the square root of the MSE.
- $\hat{\sigma}_e$: The square root of MSE serves as an estimate of the standard deviation of the error term.
- Other names for root MSE:
  - standard error of the estimate (in SPSS)
  - Residual standard error (in R)
  - $std.err.(e)$.

Outline

R^2. The Coefficient of Determination

- R^2 is the "coefficient of determination"
- R^2 has a minimum of 0 and a maximum of 1.
- R^2 mostly about "how big" the error variance is compared to the variance of x and y.

The "Proportion of Variance Explained"

- Some people write that the R^2 represents the proportion of variance in y explained by x. Where do they get that?
- The Total Sum of Squares: $TSS = \sum(y_i - \bar{y})^2$
- The Error Sum of Squares: $ESS = \sum(y_i - \hat{y}_i)^2$
- Regression Sum of Squares
  - $RSS = TSS - ESS$
  - $RSS = \sum(\hat{y}_i - \bar{y}_i)^2$

"Proportion of Variance"(cont)

- Notice

$$TSS = RSS + ESS$$

- Divide all terms by TSS and we see that the two "proportions" of variance add up to one

$$1 = \frac{RSS}{TSS} + \frac{ESS}{TSS}$$

- That's

$$1 = \textit{part accounted for by regression} + \textit{part accounted for by error} \tag{12}$$

"Proportion of Variance"(cont)

- Let the "coefficient of determination" be

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

- which is the same as
-

$$1 - \frac{ESS}{TSS}$$

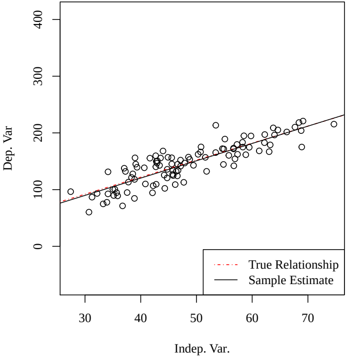
- Put that in words: R^2 is the proportion of variance left over after we take out the part contributed by random error term.
- Calculate the 'anova' table for a regression model, you'll see for yourself.

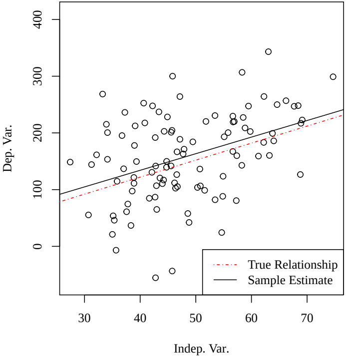
How Important is R^2

- Experienced statisticians may have rules of thumb about R^2. For example, R^2 should be bigger than 0.2 before a model is worth reporting.
- For various reasons (next slides), I think that's silly.
- Sometimes practitioners think a low R^2 is a general warning sign that "something is wrong."
- That's also mistaken: it might be there's not powerful predictive relationship to be found. We shouldn't torture the data.
- R^2 is partly dependent on the error term's variance, and we will see later that big variance -> wide confidence intervals. I often do wish error variance were smaller.

Don't Over-Emphasize R^2

- A slope is a slope is a slope, no matter how big the error variance might be. The same b's underlie both, but $R^2 = 0.70$ on left and 0.15 on right:

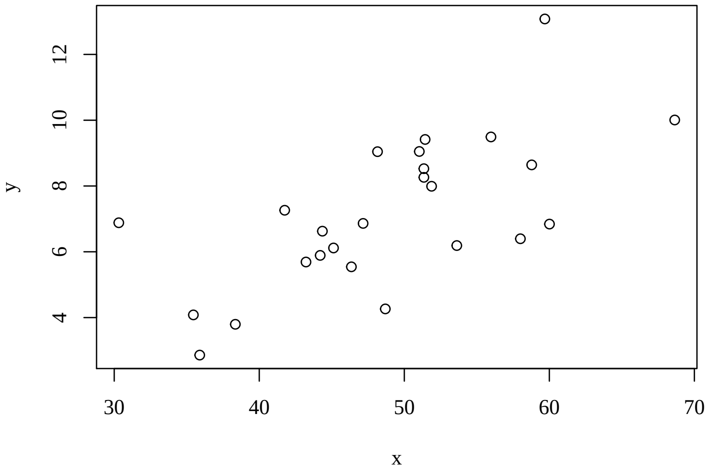




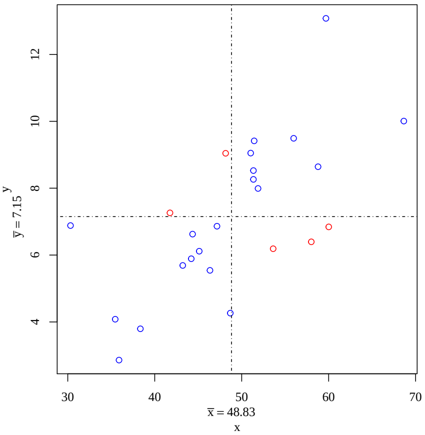
R^2continued

- The R^2 reflects 3 factors that melt together
  - The range of x
  - The size of the slope coefficient
  - The standard deviation of the error term.
- Any of those 3 culprits can make the R^2 shrink.
- Does not necessarily imply that some better regression model exists–it may just be that the process under study has inherent uncertainty.
- Careful: Wrong to compare R^2 across models with different data. (Both $Var[x_i]$ and $Var[e_i]$ can change across data sets.)

A Scatterplot: How Strongly Are These Variables Related?

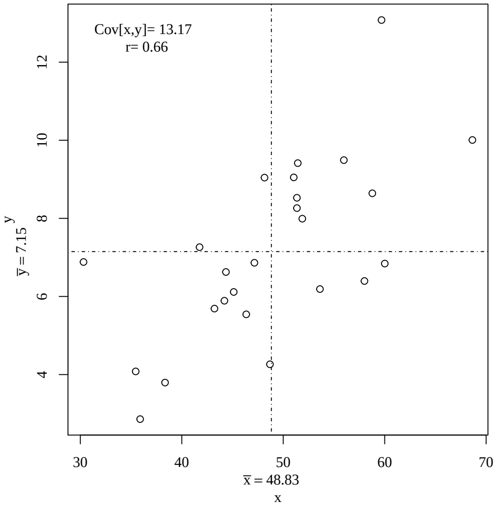


Covariance: Consider the Quadrants

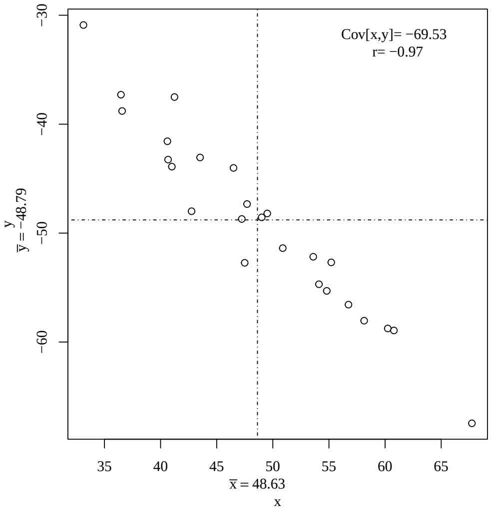


- Covariance: For each point, calculate $(x_i - \bar{x})(y_i - \bar{y})$
- Covariance: add those up, divide by N.
- blue points have positive products
- red points have negative products

How strong is this relationship?



Is this relationship stronger?



Correlation=scaled covariance

- Question: How do you know if $\widehat{Cov}[x,y]$ is "big" or "medium" or "small"
- Karl Pearson's Answer: form a correlation coefficient by scaling the covariance

$$r = \frac{\widehat{Cov}[x,y]}{\widehat{Std.Dev.}[x] \cdot \widehat{Std.Dev.}[y]} \tag{13}$$

- $r \in [-1, 1]$. That's all I know for sure about Pearson's r.

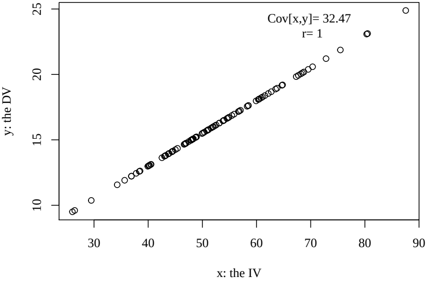
If there is One Input

- The Pearson's r squared equals the R^2 in a one-predictor regression.
- Since we already argued that R^2 has a "proportion of variance accounted for" interpretation, that means Pearson's r squared has same meaning.
- The r_{yx} (and R^2) balance Covariance against the variance of x and y.

Simulate Data For Regression

This has no "random error term" ($e_i = 0$)

- $\beta_0 = 3$
- $\beta_1 = 0.25$
- $x_i \sim N(50, 100),\ i = \{1, 2, \ldots 100\}$
- $y_i = \beta_0 + \beta_1 x_i$



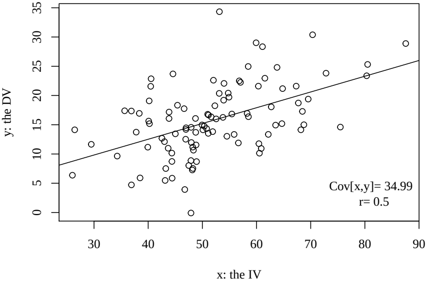
There's no "error term"

Add Some Error to y_i to adjust σ_e (and hence R^2)

- Same β_0=3, $\beta_1 = 0.25$, x_i
- $y_i = \beta_0 + \beta_1 x_i + e_i$
- $e_i \sim N(0, 5^2)$

	M1	
	Estimate	(S.E.)
(Intercept)	1.743	(2.524)
x	0.269***	(0.047)
N	100	
RMSE	5.375	
R^2	0.248	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$



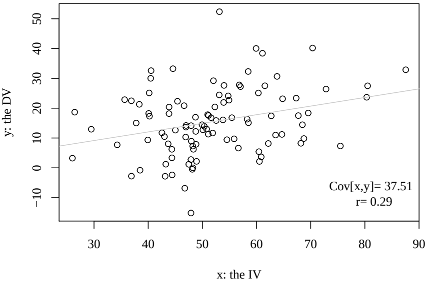
Std. Deviation of error term is 5

Tune Up Std.Dev.(e) -> Shrink the Correlation

- Same $\beta_0 = 3,\ \beta_1 = 0.25,\ x_i$
- $y_i = \beta_0 + \beta_1 x_i + e_i$
- $e_i \sim N(0, 10^2)$

	M1	
	Estimate	(S.E.)
(Intercept)	0.487	(5.047)
x	0.289**	(0.095)
N	100	
RMSE	10.749	
R^2	0.087	

$*p \le 0.05 ** \ p \le 0.01 *** p \le 0.001$



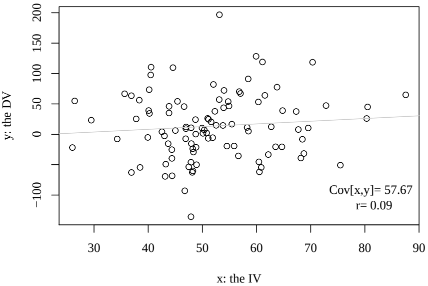
Std. Deviation of error term is 10

A Massive Std.Dev.(e) Makes R^2 Even Smaller

- Same β_0, β_1, x
- $y_i = \beta_0 + \beta_1 x_i + e_i$
- $e_i \sim N(0, 50^2)$

	M1	
	Estimate	(S.E.)
(Intercept)	-9.567	(25.237)
x	0.444	(0.474)
N	100	
RMSE	53.745	
R^2	0.009	

$*p \leq 0.05 ** \; p \leq 0.01 *** p \leq 0.001$



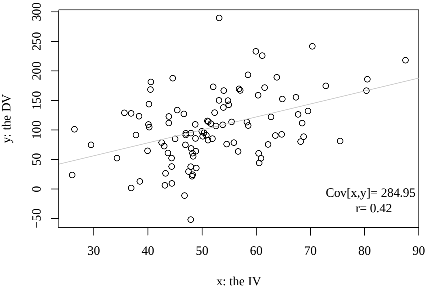
Std. Deviation of error term is 50

Leave Std.Dev.(e) Large, but Raise b_1

- Same β_0, x, and $e_i \sim N(0, 50^2)$
- Make β_1 bigger

	M1	
	Estimate	(S.E.)
(Intercept)	-9.567	(25.237)
x	2.194***	(0.474)
N	100	
RMSE	53.745	
R^2	0.179	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$



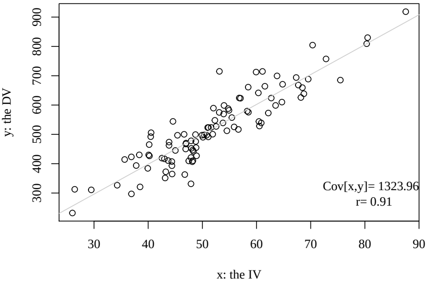
Std. Deviation of error term is 50, $\beta_1 = 2$

Make b_1 Even Larger

- Same β_0, β_1, x
- $y_i = \beta_0 + \beta_1 x_i + e_i$
- $e_i \sim N(0, 50^2)$

	M1	
	Estimate	(S.E.)
(Intercept)	-9.567	(25.237)
x	10.194***	(0.474)
N	100	
RMSE	53.745	
R^2	0.825	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$



Std. Deviation of error term is 50 and $\beta_1 = 10$

What are you Supposed to Conclude?

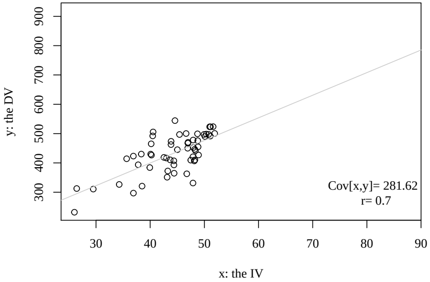
- The slope and the error variance are "balancing" each other.
- If the error variance is large, we need a steep slope to compensate and keep R^2 in the same vicinity.
- We can also fiddle with R^2 by adjusting the range of x (shown next).

A Restricted x Range Makes r Smaller

- Chopped off the top half of the x_i observations
- Wow. The effect of x on y is the same, $\beta_1 = 10$
- Smaller $Var(x) \rightarrow$ Smaller R^2 ("design" implication)

	M1	
	Estimate	(S.E.)
(Intercept)	91.217	(48.138)
x2	7.709***	(1.080)
N	56	
RMSE	48.419	
R^2	0.485	

$*p \leq 0.05 ** p \leq 0.01 *** p \leq 0.001$



Std. Deviation of error term is 50 and $\beta_1 = 10$

Section Summary

- Correlation depends on several components, $Var(x_i)$, b_1, and $Var(e_i)$.
- The "correlation coefficient" is not a "parameter." It is a description or a 'weighted summary' of the effect of parameters on the data.
- Goldberger (1991, p.177) puts it the following way: "Nothing in the CR (Classical Regression) model requires that R^2 be high. Hence, a high R^2 is not evidence in favor of the model, and a low R^2 is not evidence against it."
- Nevertheless, R^2 can be a persuasive tool because many people think a model is "wrong" if the R^2 does not meet some subjective standard.

Outline

SMW: Use Calculus to Minimize $S(\hat{\beta}_0, \hat{\beta}_1)$

- Must find the minimum S, which is shaped like a bowl
- Find combination of $(\hat{\beta}_0\ \hat{\beta}_1)$ where the function is “flat”, at bottom of bowl
- First Order Conditions:

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = 0 \qquad (14)$$

and

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = 0 \qquad (15)$$

Sketch something here:

SMW: First Order Condition for $\hat{\beta}_0$:

$$\begin{aligned}\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} &= -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i) = 0 \\ &= \sum y_i - \sum \hat{\beta}_0 - \sum \hat{\beta}_1 \cdot x_i = 0 \\ &= \sum y_i - N \cdot \hat{\beta}_0 - \hat{\beta}_1 \cdot \sum x_i = 0 \end{aligned} \tag{16}$$

SMW: First Order Condition for $\hat{\beta}_1$:

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i) x_i = 0$$

$$= \sum y_i - \sum \hat{\beta}_0 \cdot x_i - \sum \hat{\beta}_1 \cdot x_i^2 = 0 \tag{17}$$

SMW: Normal Equations.

Equations 16 and 17 can be re-arranged as the so-called "normal equations".

$$\sum y_i = N\hat{\beta}_0 + \left(\sum x_i\right)\hat{\beta}_1$$

and

$$\sum x_i y_i = \left(\sum x_i\right)\hat{\beta}_0 + \left(\sum x_i^2\right)\hat{\beta}_1$$

SMW: Note that is a LINEAR Matrix Equation

$$\begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \tag{18}$$

Refer to the coefficient estimates as $\hat{\beta}$:

$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$,

SMW: The Solution

- The "sum of squares minimizing" estimate vector is

$$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T y \tag{19}$$

- Definition: X is predictor "design matrix", $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{N-1} \\ 1 & x_N \end{bmatrix}$

- And $y = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_{N-1} \\ y_N \end{bmatrix}$