Lecture 17

Statistics---the linear model

A little bit of review:

After establishing a foundation in probability, we proceeded to estimation of unknown parameters. (We talked about criteria for assessing them as well as where they might come from.) Most, if not all, of that foundational discussion was focused on estimating parameters of a univariate distribution (like the mean or the variance or some other parameter that characterizes it). So much of what we care about in social science (and many other settings as well) involves joint distributions, though.

Statistics---the linear model

A little bit of review:

Esther's discussion of causality was the beginning of (and a special case, really) of our consideration of the joint distribution of variables of interest and how we will estimate parameters of these joint distributions. You can think of much of what she did as considering the joint distribution of two variables where one was simply a coin flip (H: treatment, T: control) and the other was the outcome of interest (e.g., infant mortality, or website effectiveness).

Statistics---the linear model

A little bit of review:

And, in fact, we were mostly concerned with the *conditional* distribution of the outcome variable conditional on the coin flip. We can (and did) think of the treatment and control groups being two separate populations, and we were interested in, say, testing whether their means were equal. We can also think about having one population and a joint distribution of those two random variables on that population.

Statistics---the linear model

A little bit of review:

What if, instead of a coin flip, the second random variable is continuous? It can take on a whole range of values. How do we analyze the conditional distribution of our outcome variable conditional on something like a continuous random variable? How do we estimate the parameters of that conditional distribution?

Statistics---the linear model

A little bit of review:

What if, instead of a coin flip, the second random variable is continuous? It can take on a whole range of values. How do we analyze the conditional distribution of our outcome variable conditional on something like a continuous random variable? How do we estimate the parameters of that conditional distribution?

The workhorse model we use is the linear model and the way we estimate the parameters is linear regression.

Statistics---the linear model

Why do we care about joint distributions and estimating the parameters associated with them?

---prediction

---determining causality

---just understanding the world better

Statistics---the linear model

Why do we care about joint distributions and estimating the parameters associated with them?

---prediction

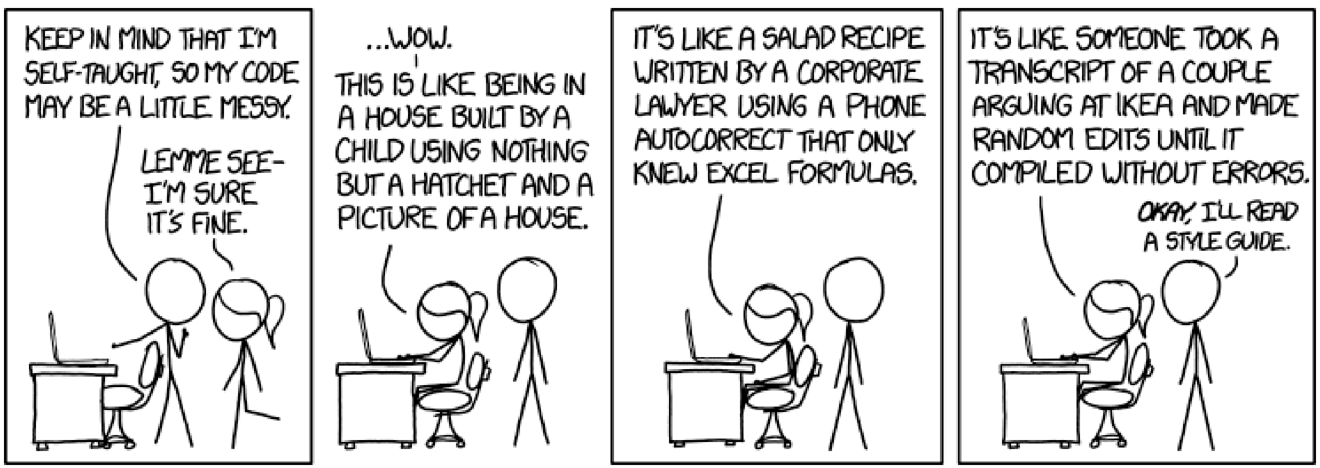
If I am the type of person who reads xkcd, am I also the type of person who is likely to click on an ad for a t-shirt bearing the Russian cover design of Moby Dick?

Statistics---the linear model

Why do we care about joint distributions and estimating the parameters associated with them?

---prediction

If I am the type of person who reads xkcd, am I also the type of person who is likely to click on an ad for a t-shirt bearing the Russian cover design of Moby Dick?



my favorite xkcd

t-shirt from Out of Print

Moby-Dick: Russian Edition

$ 28

Statistics---the linear model

Why do we care about joint distributions and estimating the parameters associated with them?

---determining causality

If I give my dog a treat every time he does not bark at another dog walking by our house, will he stop barking at other dogs?

Statistics---the linear model

Why do we care about joint distributions and estimating the parameters associated with them?

---determining causality

If I give my dog a treat every time he does not bark at another dog walking by our house, will he stop barking at other dogs?

Statistics---the linear model

Why do we care about joint distributions and estimating the parameters associated with them?

---just understanding the world better

Are people only influenced by price, quality, characteristics, and expected weather when they purchase a convertible, or are they also influenced by the weather on that particular day?

Statistics---the linear model

Why do we care about joint distributions and estimating the parameters associated with them?

---just understanding the world better

The Psychological Effect of Weather on Car Purchases*

Meghan R. Busse, Devin G. Pope, Jaren C. Pope and Jorge Silva-Risso

+ Author Affiliations

Abstract

When buying durable goods, consumers must forecast how much utility they will derive from future consumption, including consumption in different states of the world. This can be complicated for consumers because making intertemporal evaluations may expose them to a variety of psychological biases such as present bias, projection bias, and salience effects. We investigate whether consumers are affected by such intertemporal biases when they purchase automobiles. Using data for more than 40 million vehicle transactions, we explore the impact of weather on purchasing decisions. We find that the choice to purchase a convertible or a four-wheel-drive is highly dependent on the weather at the time of purchase in a way that is inconsistent with classical utility theory. We consider a range of rational explanations for the empirical effects we find, but none can explain fully the effects we estimate. We then discuss and explore projection bias and salience as two primary psychological mechanisms that are consistent with our results. *JEL* Codes: D03; D12.

Are people only influenced by price, quality, characteristics, and expected weather when they purchase a convertible, or are they also influenced by the weather on that particular day?

QJE, 2014

Statistics---the linear model

Why do we care about joint distributions and estimating the parameters associated with them?

---just understanding the world better

The Psychological Effect of Weather on Car Purchases*

Meghan R. Busse, Devin G. Pope, Jaren C. Pope and Jorge Silva-Risso

+ Author Affiliations

Abstract

When buying durable goods, consumers must forecast how much utility they will derive from future consumption, including consumption in different states of the world. This can be complicated for consumers because making intertemporal evaluations may expose them to a variety of psychological biases such as present bias, projection bias, and salience effects. We investigate whether consumers are affected by such intertemporal biases when they purchase automobiles. Using data for more than 40 million vehicle transactions, we explore the impact of weather on purchasing decisions. We find that the choice to purchase a convertible or a four-wheel-drive is highly dependent on the weather at the time of purchase in a way that is inconsistent with classical utility theory. We consider a range of rational explanations for the empirical effects we find, but none can explain fully the effects we estimate. We then discuss and explore projection bias and salience as two primary psychological mechanisms that are consistent with our results. *JEL* Codes: D03; D12.

Are people only influenced by price, quality, characteristics, and expected weather when they purchase a convertible, or are they also influenced by the weather on that particular day?

QJE, 2014

Statistics---the linear model

In each of those examples, there were two or more random variables, jointly distributed, and we would like to know characteristics of their joint distribution in order to answer the questions.

Statistics---the linear model, bivariate style

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

random variables (on which we have repeated observations)

Statistics---the linear model, bivariate style

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

the dependent variable (or explained variable or regressand)

Statistics---the linear model, bivariate style

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

the regressor or explanatory variable (or independent variable)

Statistics---the linear model, bivariate style

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

unobserved random variable, the error

Statistics---the linear model, bivariate style

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

parameters to be estimated, the regression coefficients

Statistics---the linear model, bivariate style

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

parameters to be estimated

This model allows us to consider the mean of a random variable Y as a function of another (random) variable X. If we obtain estimates for β_0 and β_1, we than have an estimated conditional mean function for Y.

Statistics---the linear model, bivariate style

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Add basic assumptions to get classical linear regression model:

i) X_i, ε_i uncorrelated

ii) identification---$(1/n)\sum_i (X_i - \bar{X})^2 > 0$

iii) zero mean---$E(\varepsilon_i) = 0$

iv) homoskedasticity---$E(\varepsilon_i^2) = \sigma^2$ for all i

v) no serial correlation---$E(\varepsilon_i \varepsilon_j) = 0$ if $i \neq j$

Statistics---the linear model

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

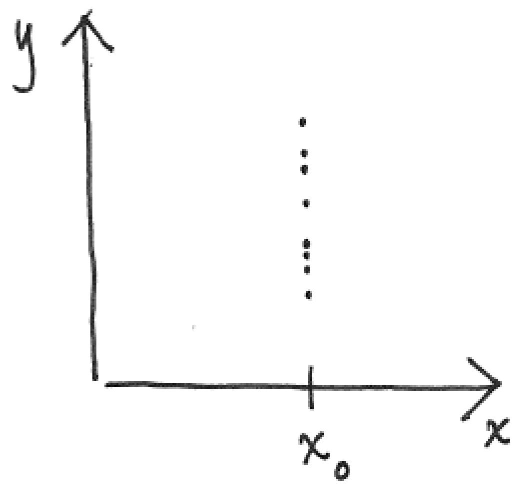
Notes:

We sometimes impose an alternative assumption to i) for our convenience: X_i are fixed in repeated samples, or nonstochastic.

Assumptions iii)-v) could be subsumed under a stronger assumption--- ε_i i.i.d. $N(0, \sigma^2)$.

Statistics---the linear model

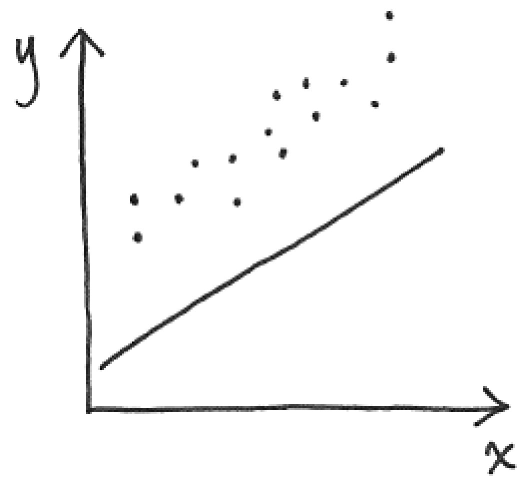
ii) identification---$(1/n)\Sigma_i(X_i-\bar{X})^2 > 0$



We rule out a case like this because it doesn't give us the variation in X that we need to identify the mean of Y conditional on X.

Statistics---the linear model

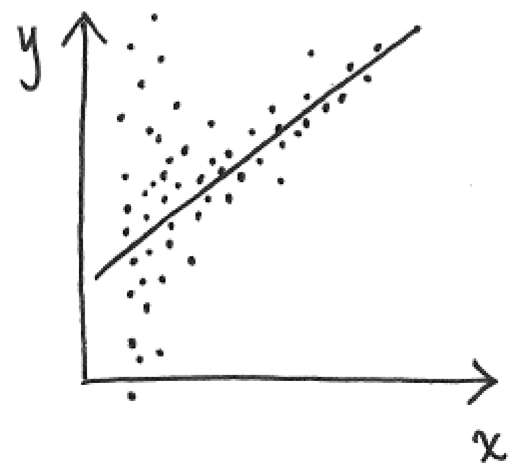
iii) zero mean---$E(\varepsilon_i) = 0$



We rule out something like this, but we don't have any information that would help us figure out whether the mean was non-zero and the intercept was just different.

Statistics---the linear model

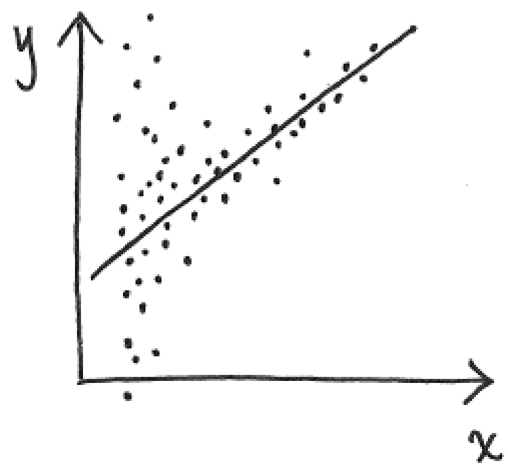
iv) homoskedasticity---$E(\varepsilon_i^2) = \sigma^2$ for all i



This is a picture of what heteroskedasticity might look like. We assume for now that we don't have it.

Statistics---the linear model

iv) homoskedasticity---$E(\varepsilon_i^2) = \sigma^2$ for all i



This is a picture of what heteroskedasticity might look like.
We assume for now that we don't have it.

Right about now you're thinking, "what is the etymology of 'homo/heteroskedasticity,' and is she even spelling it right?" (My autocorrect keeps trying to replace k with c.)

Econometrica, Vol. 53, No. 2 (March, 1985)

MISCELLANEA

ON HETEROS*EDASTICITY

BY J. HUSTON MCCULLOCH[1]

THE MOST PRESSING ISSUE in econometric orthography today is whether heteros*edasticity should be spelled with a k or with a c. Heteroskedasticity is used in their texts by Dhrymes (1970), Goldberger (1964), Intriligator (1978), Kmenta (1971) and Valavanis (1959), while heteroscedasticity is preferred by Champernowne (1969), Chow (1983), Goldfeld and Quandt (1972), Johnston (1963), Maddala (1979), Malinvaud (1970), and Theil (1971).[2]

Our word is a modern coinage, derived from the two Greek roots hetero-(ἑτερο-), meaning "other" or "different," and skedannumi (σκεδάννυμι), meaning "to scatter." The letter in question is therefore the transliteration of Greek kappa (κ).

In scientific words which scholars have lifted directly from Greek into English, the letter kappa is always transliterated as k. Examples are skeptic (σκεπτικός) and skeleton (σκελετός).

Greek kappa does sometimes make its way into English as c, but only in common words which entered English through French and old scientific words that entered through Latin. Examples are sceptre (σκῆπτρον), scene (σκηνή) and cyclic (κυκλικός). Kappa becomes c in French or Latin, simply because k is not used in these languages except to spell foreign proper names. When such a c is followed by e, i, or y, however, it is always sibillant. The only way a kappa taken into French can retain its "k" sound before one of these vowels is in the rare event that it becomes "qu" (as in *squelette*).

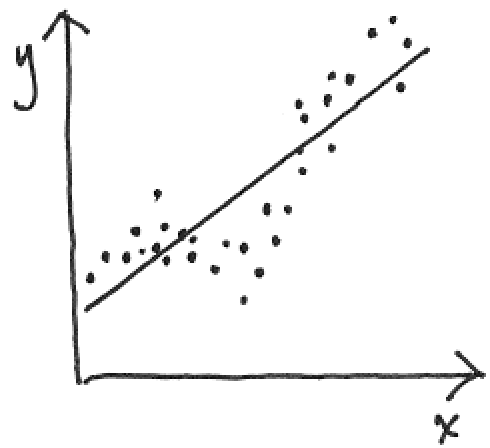
In English as in French and Latin, c before e is always soft.[3] Examples include ceiling, celerey, ceremony, cease, cedar, celestial, celibacy, cell, cement, cent, center, necessary, scent, etc., any of which would sound ridiculous with a hard c.

If heteros*edasticity were spelled with a c, it would thus have had to have entered the English language either in 1066 with the Norman invaders or else in the middle ages from Latin, neither of which was the case. Furthermore, it would have to be pronounced "heterossedasticity," which it is not.

Heteroskedasticity is therefore the proper English spelling.[4]

Statistics---the linear model

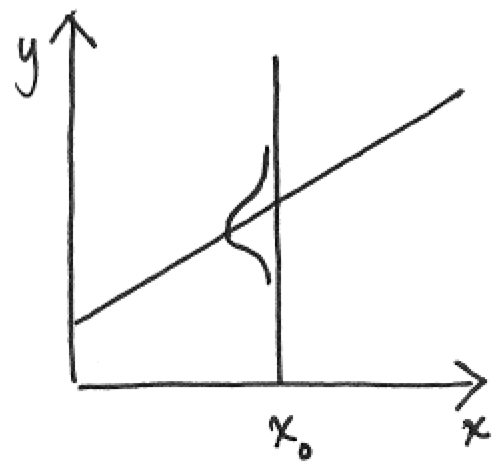
v) no serial correlation---$E(\varepsilon_i \varepsilon_j) = 0$ if $i \neq j$



This is a picture of what positive serial correlation might look like. We assume for now that we don't have it.

Statistics---the linear model

Assumptions iii)-v) could be subsumed under a stronger assumption--- ε_i i.i.d. $N(0,\sigma^2)$.



Statistics---the linear model

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Properties of model:

$E(Y_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 X_i + E(\varepsilon_i) =$
$\beta_0 + \beta_1 X_i$

$Var(Y_i) = E((Y_i - E(Y_i))^2) = E((\beta_0 + \beta_1 X_i + \varepsilon_i - \beta_0 -$
$\beta_1)^2) = E(\varepsilon_i^2) = \sigma^2$

$Cov(Y_i, Y_j) = 0$, $i \neq j$ (can show using properties of ε_i)

Statistics---the linear model

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Properties of model:

$E(Y_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 X_i + E(\varepsilon_i) =$
$\beta_0 + \beta_1 X_i$ The βs are parameters in the conditional mean function.

$Var(Y_i) = E((Y_i - E(Y_i))^2) = E((\beta_0 + \beta_1 X_i + \varepsilon_i - \beta_0 -$
$\beta_1)^2) = E(\varepsilon_i^2) = \sigma^2$

$Cov(Y_i, Y_j) = 0,\ i \neq j$ (can show using properties of ε_i)

Statistics---the linear model

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

How do we find estimates for β_0 and β_1?

---least squares: $\min_\beta \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2$

---least absolute deviations: $\min_\beta \sum_i |Y_i - \beta_0 - \beta_1 X_i|$

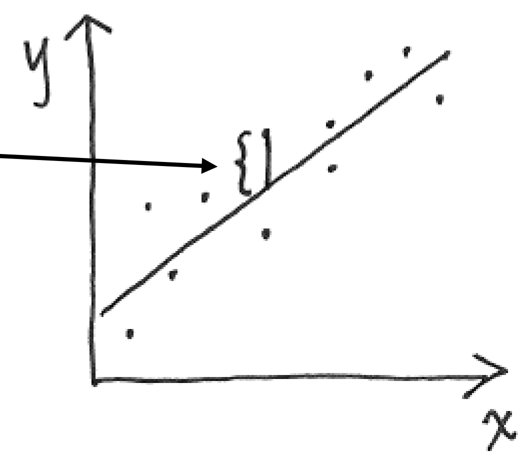
---reverse least squares: $\min_\beta \sum_i (X_i - \beta_0/\beta_1 - Y_i/\beta_1)^2$

Statistics---the linear model

---least squares: $\min_\beta \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2$

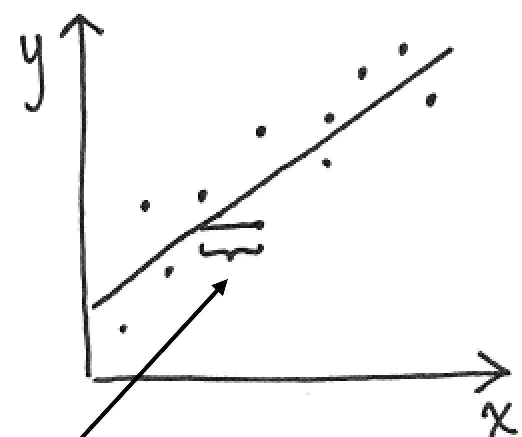
---least absolute deviations: $\min_\beta \sum_i |Y_i - \beta_0 - \beta_1 X_i|$

We minimize the sum of squares or sum of absolute values of these things.



Statistics---the linear model

---reverse least squares: $\min_\beta \sum_i (X_i - \beta_0/\beta_1 - Y_i/\beta_1)^2$



We minimize the sum of squares of these things.

Statistics---the linear model

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

We'll focus on least squares (sometimes called "ordinary least squares," or OLS). Why? Under the assumptions of the Classical Linear Regression Model, OLS provides the minimum variance (most efficient) unbiased estimator of β_0 and β_1, it is the MLE under normality of errors, and the estimates are consistent and asymptotically normal.

Statistics---the linear model

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Do we have to do a numerical minimization every time we want to solve for the least squares estimates?

No, we have lovely, closed-form solutions:

$$\hat{\beta}_1 = \{(1/n)\Sigma(X_i - \overline{X})(Y_i - \overline{Y})\}/\{(1/n)\Sigma(X_i - \overline{X})^2\}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Statistics---the linear model

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Do we have to do a numerical minimization every time we want to solve for the least squares estimates?

No, we have lovely, closed-form solutions:

$$\hat{\beta}_1 = \{(1/n)\Sigma(X_i - \overline{X})(Y_i - \overline{Y})\}/\{(1/n)\Sigma(X_i - \overline{X})^2\}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

How do we get these? Pages of tedious calculations, up on the website for your viewing pleasure.

Statistics---the linear model

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Do we have to do a numerical minimization every time we want to solve for the least squares estimates?

No, we have lovely, closed-form solutions:

$$\hat{\beta}_1 = \{(1/n)\Sigma(X_i - \overline{X})(Y_i - \overline{Y})\}/\{(1/n)\Sigma(X_i - \overline{X})^2\}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

I don't want you to get the idea that OLS estimators are horrible, complicated things. They are very elegant and intuitive, but this summation-based notation is not up to the task.

Statistics---the linear model

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

And they could be lovelier still if we weren't too afraid of using matrix notation . . .

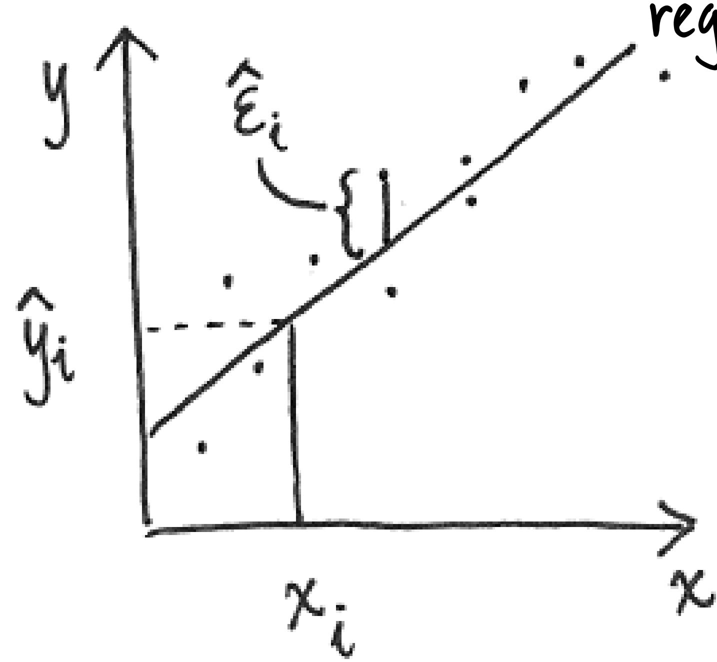
Statistics---the linear model

A couple of important definitions:

residual--- $\hat{\varepsilon}_i = y_i - \hat{y}_i$

regression line, or fitted line --- $\hat{\beta}_0 + \hat{\beta}_1 x$

fitted value---
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$



Statistics---the linear model

What do we always ask when we learn about a new estimator (and why do we ask it)?

Statistics---the linear model

What do we always ask when we learn about a new estimator (and why do we ask it)? We want to know how is it distributed (because we cannot perform inference, like creating confidence intervals and running hypothesis tests, unless we know something about its distribution).

Statistics---the linear model

What do we always ask when we learn about a new estimator (and why do we ask it)? We want to know how is it distributed (because we cannot perform inference, like creating confidence intervals and running hypothesis tests, unless we know something about its distribution).

Let $\bar{x} = \frac{1}{n}\sum x_i$ and $\hat{\sigma}_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$ (for convenience).

Statistics---the linear model

What do we always ask when we learn about a new estimator (and why do we ask it)? We want to know how is it distributed (because we cannot perform inference, like creating confidence intervals and running hypothesis tests, unless we know something about its distribution).

Let $\bar{x} = \frac{1}{n}\sum x_i$ and $\hat{\sigma}_x^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$.

	mean	variance	covariance
$\hat{\beta}_0$	β_0	$\sigma^2 \bar{x}^2 / n\hat{\sigma}_x^2 + \sigma^2/n$	$-\sigma^2 \bar{x} / n\hat{\sigma}_x^2$
$\hat{\beta}_1$	β_1	$\sigma^2 / n\hat{\sigma}_x^2$	

Statistics---the linear model

What do we always ask when we learn about a new estimator (and why do we ask it)? We want to know how is it distributed (because we cannot perform inference, like creating confidence intervals and running hypothesis tests, unless we know something about its distribution).

Let $\bar{x} = \frac{1}{n}\sum x_i$ and $\hat{\sigma}_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$

How do we get these? Pages of tedious calculations, up on the website for your viewing pleasure.

	mean	variance	covariance
$\hat{\beta}_0$	β_0	$\sigma^2\bar{x}^2/n\hat{\sigma}_x^2 + \sigma^2/n$	$-\sigma^2\bar{x}/n\hat{\sigma}_x^2$
$\hat{\beta}_1$	β_1	$\sigma^2/n\hat{\sigma}_x^2$	

Statistics---the linear model

What do we always ask when we learn about a new estimator (and why do we ask it)? We want to know how is it distributed (because we cannot perform inference, like creating confidence intervals and running hypothesis tests, unless we know something about its distribution).

Let $\bar{x} = \frac{1}{n}\sum x_i$ and $\hat{\sigma}_x^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$

These you knew because I told you that OLS estimates were unbiased.

	mean	variance	covariance
$\hat{\beta}_0$	β_0	$\sigma^2\bar{x}^2/n\hat{\sigma}_x^2 + \sigma^2/n$	$-\sigma^2\bar{x}/n\hat{\sigma}_x^2$
$\hat{\beta}_1$	β_1	$\sigma^2/n\hat{\sigma}_x^2$	

Statistics---the linear model

	mean	variance	covariance
$\hat{\beta}_0$	β_0	$\sigma^2 \bar{x}^2 / n\hat{\sigma}_x^2 + \sigma^2/n$	$-\sigma^2 \bar{x} / n\hat{\sigma}_x^2$
$\hat{\beta}_1$	β_1	$\sigma^2 / n\hat{\sigma}_x^2$	

Some comparative statics:

---A larger σ^2 means larger Var($\hat{\beta}$)

---A larger $\hat{\sigma}_x^2$ means smaller Var($\hat{\beta}$)

---A larger n means smaller Var($\hat{\beta}$)

---If $\bar{x} > 0$, Cov(β_0, β_1) < 0

Statistics---the linear model

	mean	variance	covariance
$\hat\beta_0$	β_0	$\sigma^2\bar{X}^2/n\hat\sigma_x^2 + \sigma^2/n$	$-\sigma^2\bar{X}/n\hat\sigma_x^2$
$\hat\beta_1$	β_1	$\sigma^2/n\hat\sigma_x^2$	

Some comparative statics:

the vector of parameters

---A larger σ^2 means larger $\mathrm{Var}(\hat\beta)$

---A larger $\hat\sigma_x^2$ means smaller $\mathrm{Var}(\hat\beta)$

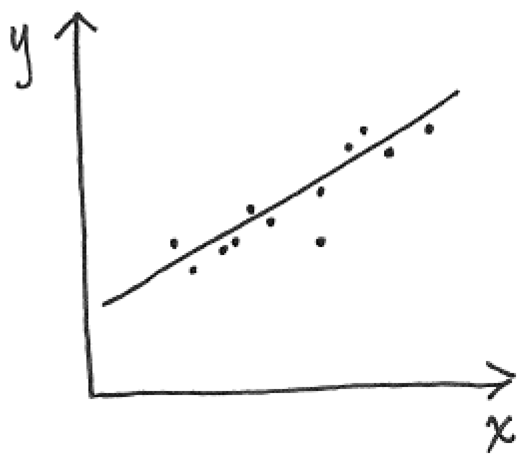
---A larger n means smaller $\mathrm{Var}(\hat\beta)$

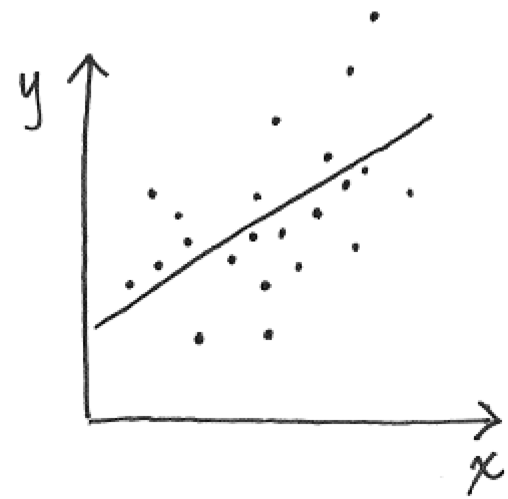
---If $\bar{X} > 0$, $\mathrm{Cov}(\beta_0, \beta_1) < 0$

Statistics---the linear model

---A larger σ^2 means larger $Var(\hat{\beta})$

variance of the error



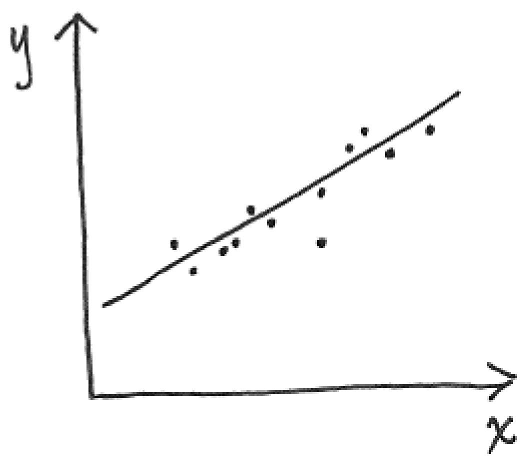


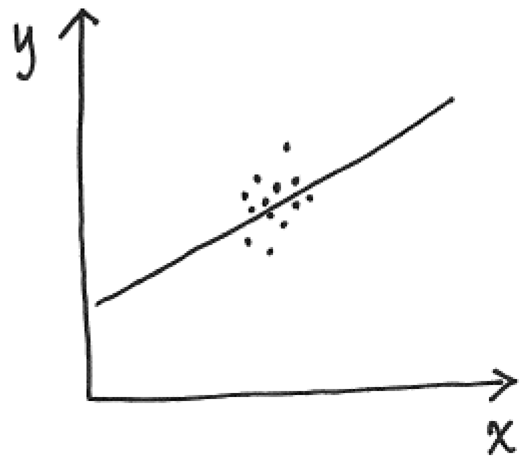
less sure of our estimates
in this case---higher variance

Statistics---the linear model

---A larger $\hat{\sigma}_x^2$ means smaller $Var(\hat{\beta})$

how much variation we have in our explanatory variable





less sure of our estimates
in this case---higher variance

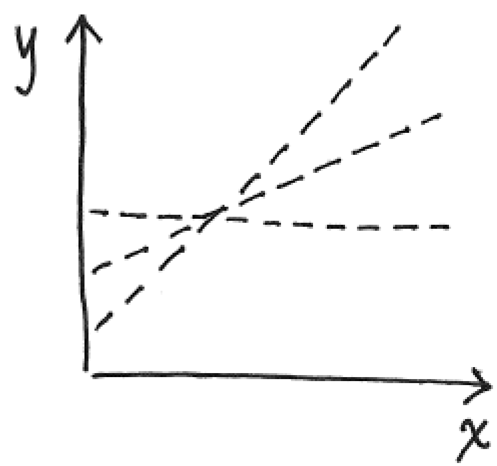
Statistics---the linear model

---A larger n means smaller Var($\hat{\beta}$)

I won't draw a picture, but we'll just note that this comparative static follows from consistency of $\hat{\beta}$.

Statistics---the linear model

---If $\bar{x} > 0$, $Cov(\beta_0, \beta_1) < 0$



a mechanical relationship
between the two estimates

Statistics---the linear model

One step further: If we use the stronger assumption that the errors are i.i.d. $N(0,\sigma^2)$, we obtain the result that $\hat{\beta}_0$ and $\hat{\beta}_1$ will also have normal distributions.

Statistics---the linear model

Note that the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of σ^2. But we often don't know σ^2. So we estimate it.

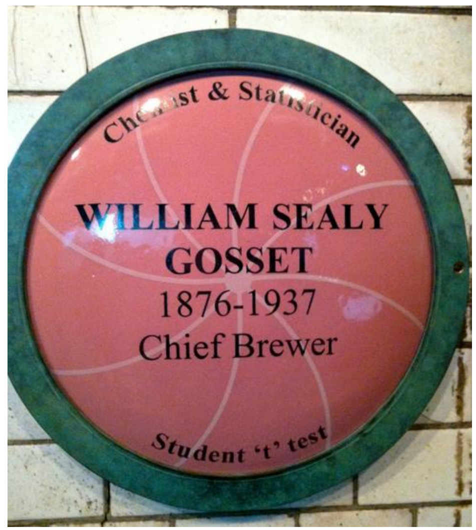
Let's use $\hat{\sigma}^2 = \frac{1}{n-2}\sum \hat{\varepsilon}_i^2$ because it's unbiased for σ^2. (Why the -2 in the denominator? Because we're estimating two parameters, β_0 and β_1, and it turns out that's what we need for $\hat{\sigma}^2$ to be unbiased.)

Statistics---the linear model

What happened when we were doing univariate inference and we replaced an unknown variance with an estimate of the variance?

Statistics---the linear model

What happened when we were doing univariate inference and we replaced an unknown variance with an estimate of the variance?



Same thing is going to happen here.

Statistics---the linear model

Now that we have all of the pieces (model, estimators, information about the distribution of estimators, etc.), we could proceed with inference, but we're going to put that off for a little while. For now, let's take a quick detour: analysis of variance.

Statistics---the linear model

We want some way to indicate how closely associated X and Y are, or how much of Y's variation is "explained" by X's variation. We perform an analysis of variance and that leads us to a measure of goodness-of-fit.

Statistics---the linear model

We want some way to indicate how closely associated X and Y are, or how much of Y's variation is "explained" by X's variation. We perform an analysis of variance and that leads us to a measure of goodness-of-fit.

Let's start by defining the sum of squared residuals, SSR.

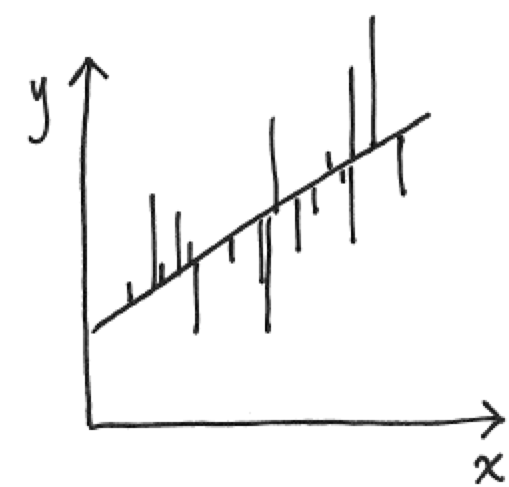
$$SSR = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_i (\hat{\varepsilon}_i)^2$$

Statistics---the linear model

We want some way to indicate how closely associated X and Y are, or how much of Y's variation is "explained" by X's variation. We perform an analysis of variance and that leads us to a measure of goodness-of-fit.

Let's start by defining the sum of squared residuals, SSR.

$$SSR = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_i (\hat{\varepsilon}_i)^2$$



Statistics---the linear model

$$SSR = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_i (\hat{\varepsilon}_i)^2$$

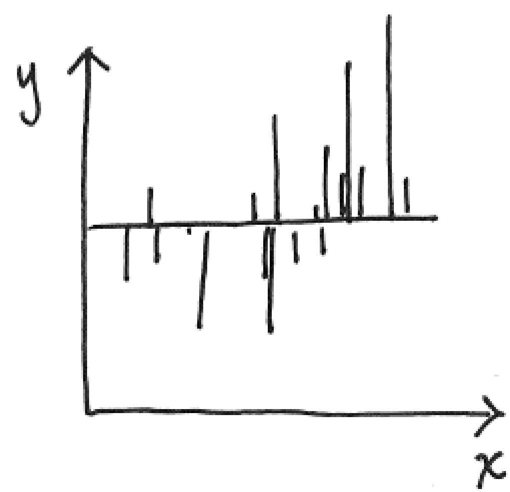
This is, in some sense, a measure of goodness-of-fit, but it is not unit-free, which is inconvenient. If we divide by the total sum of squares, that gives us a unit-free measure:

$$SST = \sum_i (Y_i - \bar{Y})^2$$

Statistics---the linear model

$$SSR = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_i (\hat{\varepsilon}_i)^2$$

This is, in some sense, a measure of goodness-of-fit, but it is not unit-free, which is inconvenient. If we divide by the total sum of squares, that gives us a unit-free measure:



$$SST = \sum_i (Y_i - \bar{Y})^2$$

Statistics---the linear model

So we have

$$SSR = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_i (\hat{\varepsilon}_i)^2$$

$$SST = \sum_i (Y_i - \bar{Y})^2$$

Note that

$0 <= SSR/SST <= 1$ Why?

Statistics---the linear model

So we have

$$SSR = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_i (\hat{\varepsilon}_i)^2$$

$$SST = \sum_i (Y_i - \bar{Y})^2$$

Note that

0 <= SSR/SST <= 1 Why? Because both of these values are non-negative, by construction, and the fact that the regression line is the "least squares" line ensures that SSR <= SST.

Statistics---the linear model

I guess we wanted a measure of fit that had larger values when the fit was better, or we explained more, so we defined

$R^2 = 1 - SSR/SST$.

It turns out that SST can be decomposed into two terms, SSR and the model sum of squares, SSM.

$$SSM = \sum_i (\hat{Y}_i - \bar{Y})^2$$

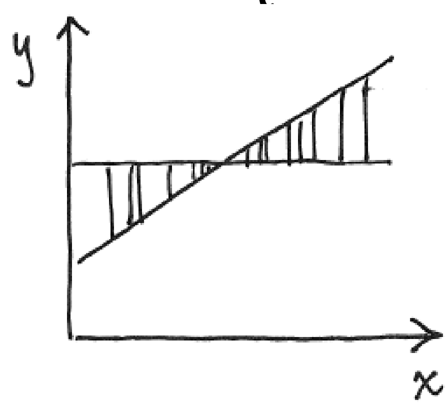
Statistics---the linear model

I guess we wanted a measure of fit that had larger values when the fit was better, or we explained more, so we defined

$R^2 = 1 - SSR/SST$.

It turns out that SST can be decomposed into two terms, SSR and the model sum of squares, SSM.

$$SSM = \sum_i (\hat{Y}_i - \bar{Y})^2$$



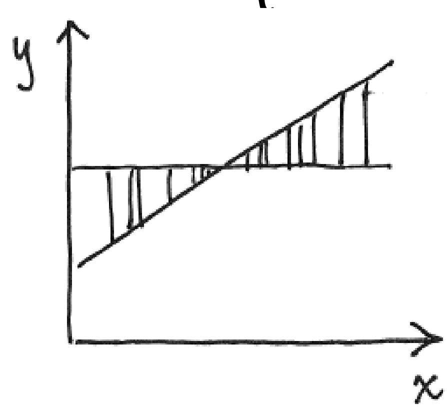
Statistics---the linear model

I guess we wanted a measure of fit that had larger values when the fit was better, or we explained more, so we defined

$R^2 = 1 - SSR/SST.$

It turns out that SST can be decomposed into two terms, SSR and the model sum of squares, SSM.

$$SSM = \sum_i (\hat{Y}_i - \bar{Y})^2$$



Cross term goes away because of how $\hat{\beta}$ is chosen.

Statistics---the linear model

In bivariate regression, $R^2 = r_{XY}^2$, the sample correlation coefficient for X and Y. R^2 is a more general formulation, though, and is defined for linear models with more than one explanatory variable.

In addition to using R^2 as a basic measure of goodness-of-fit, we can also use it as the basis of a test of the hypothesis that $\beta_1 = 0$ (or $\beta_1 = \ldots = \beta_k = 0$ if we have k explanatory variables). We reject the hypothesis when $(n-2)R^2/(1-R^2)$, which has an F distribution under the null, is large.

Statistics---the linear model

Let's talk about a few practical issues, introduce multiple regression (with matrix notation), and then return to inference. (It's just that this summation-based notation is so clunky, we'll all be happier to see confidence intervals, t-tests, and F-tests in more elegant notation.)

Statistics---the linear model, practicalities

What does regression output look like? How do we interpret it?

Here's some output from Stata on two separate bivariate regressions:

```
.
. ***********************************************
. *****   NEW TABLE 6: Advertising Intensity *****
. ***********************************************
. /* RESULTS regressions of detail-sales ratio with revenue, revenue^2, gini */
.
. reg ds lhd3rev if dropdet==0 & ds < 0.2

      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  1,    67) =    1.90
       Model |  .000105715     1  .000105715           Prob > F      =  0.1727
    Residual |  .003728207    67  .000055645           R-squared     =  0.0276
-------------+------------------------------           Adj R-squared =  0.0131
       Total |  .003833922    68  .000056381           Root MSE      =  .00746

------------------------------------------------------------------------------
          ds |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lhd3rev |   .0006272    .000455     1.38   0.173     -.000281    .0015354
       _cons |  -.0011316    .004421    -0.26   0.799     -.009956    .0076928
------------------------------------------------------------------------------

. reg js lhd3rev if dropjrn==0 & js < 0.3

      Source |       SS       df       MS              Number of obs =      70
-------------+------------------------------           F(  1,    68) =    4.50
       Model |  .002022371     1  .002022371           Prob > F      =  0.0376
    Residual |  .030570751    68   .00044957           R-squared     =  0.0620
-------------+------------------------------           Adj R-squared =  0.0483
       Total |  .032593122    69  .000472364           Root MSE      =   .0212

------------------------------------------------------------------------------
          js |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lhd3rev |   .0027332   .0012887     2.12   0.038     .0001617    .0053047
       _cons |  -.0125051   .0125445    -1.00   0.322    -.0375373    .0125271
------------------------------------------------------------------------------

.
```

Statistics---the linear model, practicalities

What does regression output look like? How do we interpret it?

$\hat{\beta}_1$

$\hat{\beta}_0$

standard errors

```
.
. ***********************************************
. *****   NEW TABLE 6: Advertising Intensity *****
. ***********************************************
. /* RESULTS regressions of detail-sales ratio with revenue, revenue^2, gini */
.
. reg ds lhd3rev if dropdet==0 & ds < 0.2

      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  1,    67) =    1.90
       Model |  .000105715     1  .000105715           Prob > F      =  0.1727
    Residual |  .003728207    67  .000055645           R-squared     =  0.0276
-------------+------------------------------           Adj R-squared =  0.0131
       Total |  .003833922    68  .000056381           Root MSE      =  .00746

------------------------------------------------------------------------------
          ds |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lhd3rev |   .0006272    .000455     1.38   0.173     -.000281    .0015354
       _cons |  -.0011316    .004421    -0.26   0.799     -.009956    .0076928
------------------------------------------------------------------------------

. reg js lhd3rev if dropjrn==0 & js < 0.3

      Source |       SS       df       MS              Number of obs =      70
-------------+------------------------------           F(  1,    68) =    4.50
       Model |  .002022371     1  .002022371           Prob > F      =  0.0376
    Residual |  .030570751    68   .00044957           R-squared     =  0.0620
-------------+------------------------------           Adj R-squared =  0.0483
       Total |  .032593122    69  .000472364           Root MSE      =   .0212

------------------------------------------------------------------------------
          js |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lhd3rev |   .0027332   .0012887     2.12   0.038     .0001617    .0053047
       _cons |  -.0125051   .0125445    -1.00   0.322    -.0375373    .0125271
------------------------------------------------------------------------------

.
```

Statistics---the linear model, practicalities

What does regression output look like? How do we interpret it?

```
.
. **********************************************
. *****   NEW TABLE 6: Advertising Intensity *****
. **********************************************
. /* RESULTS regressions of detail-sales ratio with revenue, revenue^2, gini */
.
. reg ds lhd3rev if dropdet==0 & ds < 0.2

      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  1,    67) =    1.90
       Model |  .000105715     1  .000105715           Prob > F      =  0.1727
    Residual |  .003728207    67  .000055645           R-squared     =  0.0276
-------------+------------------------------           Adj R-squared =  0.0131
       Total |  .003833922    68  .000056381           Root MSE      =  .00746

------------------------------------------------------------------------------
          ds |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lhd3rev |   .0006272    .000455     1.38   0.173     -.000281    .0015354
       _cons |  -.0011316    .004421    -0.26   0.799    -.009956    .0076928
------------------------------------------------------------------------------

. reg js lhd3rev if dropjrn==0 & js < 0.3

      Source |       SS       df       MS              Number of obs =      70
-------------+------------------------------           F(  1,    68) =    4.50
       Model |  .002022371     1  .002022371           Prob > F      =  0.0376
    Residual |  .030570751    68   .00044957           R-squared     =  0.0620
-------------+------------------------------           Adj R-squared =  0.0483
       Total |  .032593122    69  .000472364           Root MSE      =   .0212

------------------------------------------------------------------------------
          js |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lhd3rev |   .0027332   .0012887     2.12   0.038     .0001617    .0053047
       _cons |  -.0125051   .0125445    -1.00   0.322    -.0375373    .0125271
------------------------------------------------------------------------------
```

Here are results for the F-test I briefly mentioned.

We would fail to reject the null that $\beta_1 = 0$ (for any reasonably sized test).

Statistics---the linear model, practicalities

What does regression output look like? How do we interpret it?

```
.
. ***********************************************
. *****   NEW TABLE 6: Advertising Intensity *****
. ***********************************************
. /* RESULTS regressions of detail-sales ratio with revenue, revenue^2, gini */
.
. reg ds lhd3rev if dropdet==0 & ds < 0.2

      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  1,    67) =    1.90
       Model |  .000105715     1  .000105715           Prob > F      =  0.1727
    Residual |  .003728207    67  .000055645           R-squared     =  0.0276
-------------+------------------------------           Adj R-squared =  0.0131
       Total |  .003833922    68  .000056381           Root MSE      =  .00746

------------------------------------------------------------------------------
          ds |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lhd3rev |   .0006272    .000455     1.38   0.173     -.000281    .0015354
       _cons |  -.0011316    .004421    -0.26   0.799     -.009956    .0076928
------------------------------------------------------------------------------

. reg js lhd3rev if dropjrn==0 & js < 0.3

      Source |       SS       df       MS              Number of obs =      70
-------------+------------------------------           F(  1,    68) =    4.50
       Model |  .002022371     1  .002022371           Prob > F      =  0.0376
    Residual |  .030570751    68   .00044957           R-squared     =  0.0620
-------------+------------------------------           Adj R-squared =  0.0483
       Total |  .032593122    69  .000472364           Root MSE      =   .0212

------------------------------------------------------------------------------
          js |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lhd3rev |   .0027332   .0012887     2.12   0.038     .0001617    .0053047
       _cons |  -.0125051   .0125445    -1.00   0.322    -.0375373    .0125271
------------------------------------------------------------------------------

.
```

For this one, we would reject the null that $\beta_1 = 0$ for a 5% test, but not a 1% test.

Statistics---the linear model, practicalities

What does regression output look like? How do we interpret it?

```
.
. ***********************************************
. *****   NEW TABLE 6: Advertising Intensity *****
. ***********************************************
. /* RESULTS regressions of detail-sales ratio with revenue, revenue^2, gini */
.
. reg ds lhd3rev if dropdet==0 & ds < 0.2

      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  1,    67) =    1.90
       Model |  .000105715     1  .000105715           Prob > F      =  0.1727
    Residual |  .003728207    67  .000055645           R-squared     =  0.0276
-------------+------------------------------           Adj R-squared =  0.0131
       Total |  .003833922    68  .000056381           Root MSE      =  .00746

------------------------------------------------------------------------------
          ds |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lhd3rev |   .0006272    .000455     1.38   0.173     -.000281    .0015354
       _cons |  -.0011316    .004421    -0.26   0.799    -.009956    .0076928
------------------------------------------------------------------------------

. reg js lhd3rev if dropjrn==0 & js < 0.3

      Source |       SS       df       MS              Number of obs =      70
-------------+------------------------------           F(  1,    68) =    4.50
       Model |  .002022371     1  .002022371           Prob > F      =  0.0376
    Residual |  .030570751    68   .00044957           R-squared     =  0.0620
-------------+------------------------------           Adj R-squared =  0.0483
       Total |  .032593122    69  .000472364           Root MSE      =   .0212

------------------------------------------------------------------------------
          js |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lhd3rev |   .0027332   .0012887     2.12   0.038     .0001617    .0053047
       _cons |  -.0125051   .0125445    -1.00   0.322    -.0375373    .0125271
------------------------------------------------------------------------------

.
```

These are t-tests for individual coefficients. We'll get to them later.

Statistics---the linear model, practicalities

What does regression output look like? How do we interpret it?

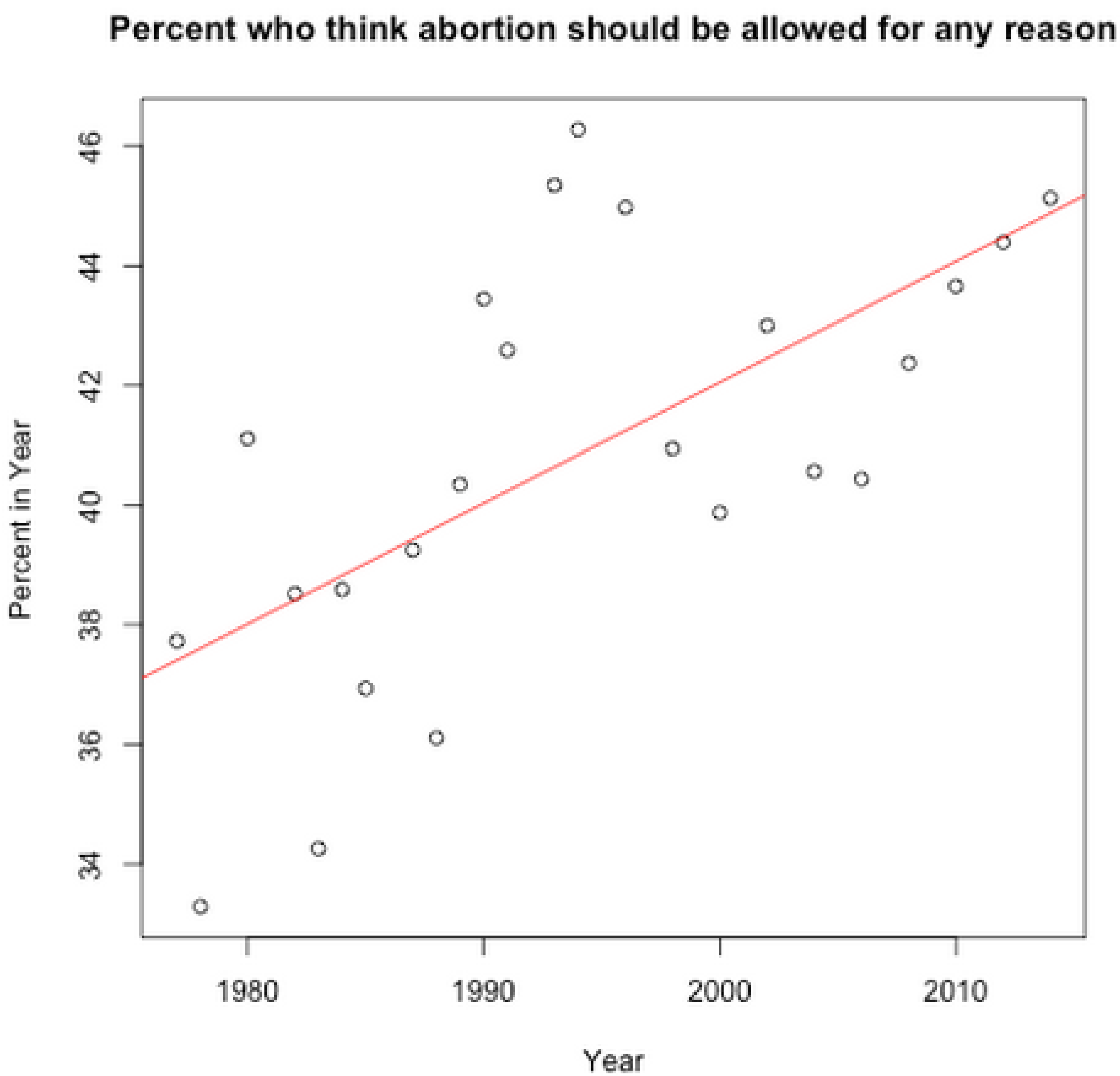
```
> fit<-lm(gss_data$any_reason~gss_data$year)
> summary(fit)

Call:
lm(formula = gss_data$any_reason ~ gss_data$year)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3595 -2.1089 -0.1308  0.9966  5.4378 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -362.02694  102.99766  -3.515 0.001953 ** 
gss_data$year    0.20204    0.05166   3.911 0.000749 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.764 on 22 degrees of freedom
Multiple R-squared:  0.4101,	Adjusted R-squared:  0.3833 
F-statistic:  15.3 on 1 and 22 DF,  p-value: 0.000749
```

$\hat{\beta}_0$

$\hat{\beta}_1$

t-tests

F-test---we would fail to reject the null that $\beta_1 = 0$ (for any reasonably sized test).



Percent who think abortion should be allowed for any reason
Percent in Year
46
44
42
40
38
36
34
1980
1990
2000
2010
Year

Statistics---the linear model, practicalities

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

How do we interpret our parameter estimates, $\hat{\beta}_1$ in particular?

$\hat{\beta}_1$ is the estimated effect on Y of a one-unit increase in X. (Precise nuances of the interpretation will depend on whether we think we have estimated a causal relationship or something else. More on that later.)

Statistics--the linear model, practicalities

```
. /*  baseball regressions  */
>
> /* this program reads in baseball.dta, the stata version of a data */
> /* file downloaded from espn.com about the 2005 mlb season.       */
> /*    team is the team city (and name)                              */
> /*    wins is the number of wins in a 162 game regular season      */
> /*    rs is total runs scored all season                            */
> /*    ra is total runs allowed all season                           */
> /*    attend is total season attendance in thousands                */
> /*    rundiff is the difference between runs scored and runs        */
> /*            allowed                                               */
```

```
. regress attend wins;

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =    9.53
       Model |  3308050.96     1  3308050.96           Prob > F      =  0.0045
    Residual |  9717640.51    28   347058.59           R-squared     =  0.2540
-------------+------------------------------           Adj R-squared =  0.2273
       Total |  13025691.5    29  449161.775           Root MSE      =  589.12

------------------------------------------------------------------------------
      attend |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wins |   31.17391   10.09733     3.09   0.005     10.49047    51.85736
       _cons |  -45.62029   824.9258    -0.06   0.956    -1735.404    1644.164
------------------------------------------------------------------------------
```

Statistics---the linear model, practicalities

```
. /*  baseball regressions  */
>
> /* this program reads in baseball.dta, the stata version of a data */
> /* file downloaded from espn.com about the 2005 mlb season.         */
> /*    team is the team city (and name)                              */
> /*    wins is the number of wins in a 162 game regular season      */
> /*    rs is total runs scored all season                            */
> /*    ra is total runs allowed all season                           */
> /*    attend is total season attendance in thousands                */
> /*    rundiff is the difference between runs scored and runs        */
> /*            allowed                                               */

. regress attend wins;

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =    9.53
       Model |  3308050.96     1  3308050.96           Prob > F      =  0.0045
    Residual |  9717640.51    28   347058.59           R-squared     =  0.2540
-------------+------------------------------           Adj R-squared =  0.2273
       Total |  13025691.5    29  449161.775           Root MSE      =  589.12

------------------------------------------------------------------------------
      attend |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wins |   31.17391   10.09733     3.09   0.005     10.49047    51.85736
       _cons |  -45.62029   824.9258    -0.06   0.956    -1735.404    1644.164
------------------------------------------------------------------------------
```

One additional win is associated with an additional 31,000 fans in attendance over the course of the season.

Statistics---the linear model, practicalities

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

What if X only takes on two values, 0 or 1? We have a special name for that type of variable, a <u>dummy</u> variable. (Sometimes we call it an <u>indicator</u> variable.)

No problem---nothing we have done here rules out any particular distribution for X or possible values of X.

Statistics---the linear model, practicalities

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

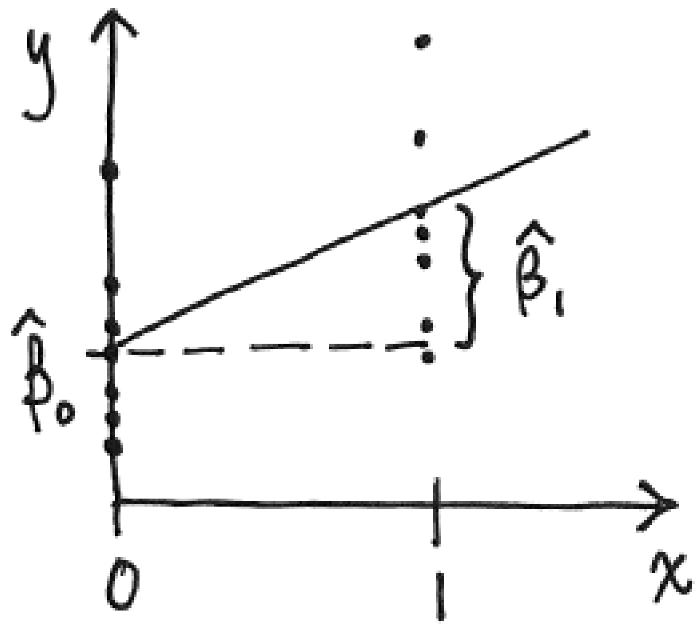
What if X only takes on two values, 0 or 1? We have a special name for that type of variable, a <u>dummy</u> variable. (Sometimes we call it an <u>indicator</u> variable.)

No problem---nothing we have done here rules out any particular distribution for X or possible values of X.

(Well, the pictures would look different.)

Statistics---the linear model, practicalities

Here's what I mean:



Statistics---the linear model, practicalities

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Dummy variables serve a number of important roles in linear models. We've (sort of) already seen one, RCTs.

Suppose we have some treatment in whose effect we are interested. We randomly assign the treatment to half of the observations and leave the other half untreated. We assign the treated observations $X = 1$ and the untreated $X = 0$.

Statistics---the linear model, practicalities

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

If we estimate the regression above, $\hat{\beta}_1$ will be the estimated effect of the treatment.

Statistics---the linear model, practicalities

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

By the way, X need not be randomly assigned half 0s and half 1s to be a dummy variable. Any characteristic that exists on some but not all observations can be represented with a dummy.

We will see other uses for dummy variables when we get to multiple regression.

Statistics---the linear model, practicalties

Linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Isn't a linear model really restrictive? What if X and Y have a relationship, but it's not linear?

---Note that the linear model is actually super flexible and can allow for all kinds of nonlinear relationships. When we get to multiple regression, we'll see some examples.

---We can do a nonparametric version, kernel regression, but there are tradeoffs, namely efficiency.

Statistics---the linear model, multivariate style

Let's consider a more general linear model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i,$$

$$i = 1, \ldots, n$$

This is a job for matrix notation!

Statistics---the linear model, multivariate style

Let's consider a more general linear model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i,$$
$$i = 1, \ldots, n$$

This is a job for matrix notation!

Let $X_i = (X_{0i}, \ldots, X_{ki})$ 1x(k+1) (row) vector (X_{0i}==1)

Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ (k+1)x1 (column) vector

Statistics---the linear model, multivariate style

So we have:

$$Y_i = X_i\beta + \varepsilon_i, \qquad i = 1, \ldots, n$$

But we can go further:

Let $Y = (Y_1, \ldots, Y_n)^T$ nx1 (column) vector

Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ nx1 (column) vector

Let $X = \begin{bmatrix} X_{01} & \ldots & X_{k1} \\ X_{02} & \ldots & X_{k2} \\ & & \\ X_{0n} & \ldots & X_{kn} \end{bmatrix}$ nx(k+1) matrix (X_{0i}==1)

Statistics---the linear model, multivariate style

So we have:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1)}{X\beta} + \underset{n\times 1}{\varepsilon}$$

Assumptions:

i) identification: $n > k+1$, X has full column rank $k+1$ (i.e., regressors are linearly independent, i.e., $X^T X$ is invertible)

ii) error behavior: $E(\varepsilon) = 0$, $E(\varepsilon\varepsilon^T)$ ($= Cov(\varepsilon)$) $= \sigma^2 I_n$ (stronger version $\varepsilon \sim N(0, \sigma^2 I_n)$)