Linear Regression

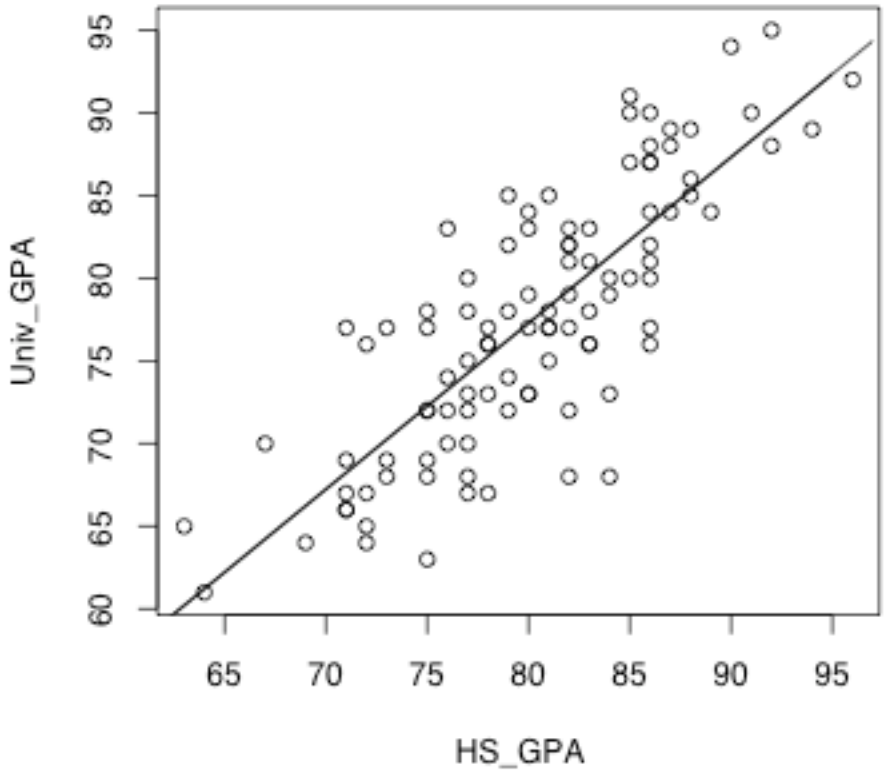
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\mathbf{X}$ is an $n \times p$ matrix of known constants
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, where $\sigma^2 > 0$ is an unknown constant.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad \widehat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

$$\mathbf{e} = (\mathbf{Y} - \widehat{\mathbf{Y}})$$

“Regression” Line



Regression Means Going Back

- Francis Galton (1822-1911) studied "Hereditary Genius" (1869) and other traits
- Heights of fathers and sons
  - Sons of the tallest fathers tended to be taller than average, but shorter than their fathers
  - Sons of the shortest fathers tended to be shorter than average, but taller than their fathers
- This kind of thing was observed for lots of traits.
- Galton was deeply concerned about "regression to mediocrity."

Measure the same thing twice, with error

$$\begin{aligned} Y_1 &= X + e_1 \\ Y_2 &= X + e_2 \end{aligned}$$

$X \sim N(\mu, \sigma_x^2)$

e_1 and e_2 independent $N(0, \sigma_e^2)$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_x^2 + \sigma_e^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2 + \sigma_e^2 \end{bmatrix} \right)$$

Conditional distribution of Y_2 given $Y_1=y_1$

$$N\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\,\rho(y_1-\mu_1), (1-\rho^2)\sigma_2^2\right)$$

$$= \quad N\left(\mu + \rho(y_1-\mu), (1-\rho^2)(\sigma_x^2+\sigma_e^2)\right)$$

So $E(Y_2|Y_1=y_1) = \mu + \rho(y_1-\mu)$,
where $\rho = \frac{\sigma_x^2}{\sigma_x^2+\sigma_e^2}$.

$$E(Y_2|Y_1 = y_1) = \mu + \rho(y_1 - \mu)$$

- If y_1 is above the mean, average y_2 will also be above the mean
- But only a fraction (rho) as far above as y_1.
- If y_1 is below the mean, average y_2 will also be below the mean
- But only a fraction (rho) as far below as y_1.

- This exactly the “regression toward the mean” that Galton observed.

Regression toward the mean

- Does not imply systematic change over time
- Is a characteristic of the bivariate normal and other joint distributions

- Can produce very misleading results, especially in the evaluation of social programs

Regression Artifact

- Measure something important, like performance in school or blood pressure.
- Select an extreme group, usually those who do worst on the baseline measure.
- Do something to help them, and measure again.

$$E(Y_2|Y_1 = y_1) = \mu + \rho(y_1 - \mu)$$

- If the treatment does **nothing**, they are expected to do worse than average, but better than they did the first time – completely artificial!

A simulation study

- Measure something twice with error: 500 observations
- Select the best 50 and the worst 50
- Do two-sided matched t-tests at alpha = 0.05

- What proportion of the time do the worst 50 show significant average improvement?
- What proportion of the time do the worst 50 show significant average deterioration?

```
> sig2x = 10; sig2e = 10; n = 500; set.seed(9999)
> X = rnorm(n,100,sqrt(sig2x))
> e1 = rnorm(n,0,sqrt(sig2e)); e2 = rnorm(n,0,sqrt(sig2e))
> Y1 = X+e1; Y2 = X+e2; D = Y2-Y1 # D measures "improvement"
> low50 = D[rank(Y1)<=50]; hi50 = D[rank(Y1)>450]
> t.test(low50)
```

```
One Sample t-test

data:  low50
t = 7.025, df = 49, p-value = 6.068e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.874234 5.177526
sample estimates:
mean of x
  4.02588
```

```
> t.test(hi50)


One Sample t-test

data:  hi50
t = -5.3417, df = 49, p-value = 2.373e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -4.208760 -1.907709
sample estimates:
mean of x
-3.058234
```

```
> t.test(low50)$p.value
[1] 6.068008e-09
> t.test(low50)$estimate
mean of x
  4.02588
> t.test(low50)$p.value<0.05 && t.test(low50)$estimate>0
[1] TRUE
```

```
> # Estimate probability of artificial difference
> M = 100 # Monte Carlo sample size
> dsig = numeric(M)
> for(i in 1:M)
+     {
+     X = rnorm(n,100,sqrt(sig2x))
+     e1 = rnorm(n,0,sqrt(sig2e)); e2 = rnorm(n,0,sqrt(sig2e))
+     Y1 = X+e1; Y2 = X+e2; D = Y2-Y1 # D measures "improvement"
+     low50 = D[rank(Y1)<=50]
+     ttest = t.test(low50)
+     dsig[i] = ttest$p.value<0.05 && ttest$estimate>0
+     }
> dsig
```

```
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[40] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[79] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Summary

- Source of the term “Regression”
- Regression artifact
  - Very serious
  - People keep re-inventing the same mistake
  - Can’t really blame the policy makers
  - At least the statistician should be able to warn them
  - The solution is random assignment
  - Taking difference from a baseline measurement may still be useful

Analysis of Variance

And testing

Analysis of Variance

- Variation to explain: **Total Sum of Squares**

$$\text{SSTO} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- Variation that is still unexplained: **Error Sum of Squares**

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

- Variation that is explained: **Regression** (or Model) **Sum of Squares**

$$\text{SSR=SSTO-SSE} = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$$

ANOVA Summary Table

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	$p-1$	SSR	$MSR = SSR/(p-1)$	$F = \frac{MSR}{MSE}$	p-value
Error	$n-p$	SSE	$MSE = SSE/(n-p)$		
Total	$n-1$	$SSTO$			

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

Proportion of variation in the dependent variable that is explained by the independent variables

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}}$$

Hypothesis Testing

- Overall F test for all the IVs at once,
- T-tests for each regression coefficient: Controlling for all the others, does that IV matter?
- Test a collection of IVs controlling for another collection,
- Most general: Testing whether sets of linear combinations of regression coefficients differ from specified constants.

Controlling for mother's education and father's education, are (any of) total family income, assessed value of home and total market value of all vehicles owned by the family related to High School GPA?

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_5 x_5$$

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$

(A false promise because of measurement error in education)

Full vs. Reduced Model

- You have 2 sets of variables, A and B
- Want to test B controlling for A
- Fit a model with both A and B: Call it the **Full Model**
- Fit a model with just A: Call it the **Reduced Model**

$$R_F^2 \geq R_R^2$$

When you add independent variables, R^2 can only go up

- By how much? Basis of F test.
- Same as testing H_0: All betas in set B (there are *d* of them) equal zero
- General H_0: **Lβ** = **h** (**L** is dxp, row rank d)

$$F = \frac{(SSR_F - SSR_R)/d}{MSE_F}$$

$$= \frac{(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{h})}{d\,MSE_F}$$

F test is based not just on change in R^2, but upon

$$a = \frac{R_F^2 - R_R^2}{1 - R_R^2}$$

Increase in explained variation expressed as a fraction of the variation that the reduced model does *not* explain.

$$F = \left(\frac{n-p}{d}\right)\left(\frac{a}{1-a}\right)$$

- For any given sample size, the bigger *a* is, the bigger *F* becomes.
- For any a ≠0, *F* increases as a function of *n.*
- So you can get a large *F* from strong results and a small sample, or from weak results and a large sample.

Can express *a* in terms of *F*

$$a = \frac{dF}{n - p + dF}$$

- Often, scientific journals just report *F*, numerator df = *d*, denominator df = *(n-p),* and a *p*-value.
- You can tell if it's significant, but how strong are the results? Now you can calculate it.
- This formula is less prone to rounding error than the one in terms of R-squared values

When you add independent variables to a model (with observational data)

- Statistical significance can appear when it was not present originally
- Statistical significance that was originally present can disappear
- Even the signs of the b coefficients can change, reversing the interpretation of how their variables are related to the dependent variable.
- This is explained by the earlier discussion of omitted variables.

A few More Points

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

- Are the x values really constants?
- Experimental versus observational data
- Omitted variables
- Measurement error in the independent variables

Recall Double Expectation

$$E\{Y\} = E\{E\{Y|X\}\}$$

E{Y} is a constant. E{Y|X} is a random variable, a function of X.

$$E\{E\{Y|X\}\} = \int E\{Y|X = x\}\, f(x)\, dx$$

Beta-hat is (conditionally) unbiased

$$E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}=\mathbf{x}\}=\boldsymbol{\beta}$$

Unbiased unconditionally, too

$$E\{\widehat{\boldsymbol{\beta}}\}=E\{E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}\}\}=E\{\boldsymbol{\beta}\}=\boldsymbol{\beta}$$

Perhaps Clearer

$$
\begin{aligned}
E\{\widehat{\boldsymbol{\beta}}\} &= E\{E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}\}\} \\
&= \int \cdots \int E\{\widehat{\boldsymbol{\beta}}|\mathbf{X} = \mathbf{x}\}\, f(\mathbf{x})\, d\mathbf{x} \\
&= \int \cdots \int \boldsymbol{\beta}\, f(\mathbf{x})\, d\mathbf{x} \\
&= \boldsymbol{\beta} \int \cdots \int f(\mathbf{x})\, d\mathbf{x} \\
&= \boldsymbol{\beta} \cdot 1 = \boldsymbol{\beta}.
\end{aligned}
$$

Conditional size α test, Critical region *A*

$$Pr\{F \in A | \mathbf{X} = \mathbf{x}\} = \alpha$$

$$\begin{aligned} Pr\{F \in A\} &= \int \cdots \int Pr\{F \in A | \mathbf{X} = \mathbf{x}\} f(\mathbf{x})\, d\mathbf{x} \\ &= \int \cdots \int \alpha f(\mathbf{x})\, d\mathbf{x} \\ &= \alpha \int \cdots \int f(\mathbf{x})\, d\mathbf{x} \\ &= \alpha \end{aligned}$$

Omitted variables

True model:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i$$

independently for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$

$$\begin{bmatrix} X_{i,1} \\ X_{i,2} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{bmatrix} \right)$$

with ϵ_i independent of $X_{i,1}$ and $X_{i,2}$.

X_2 is not observed

Since X_2 is not observed, it is absorbed

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \\ &= (\beta_0 + \beta_2 \mu_2) + \beta_1 X_{i,1} + (\beta_2 X_{i,2} - \beta_2 \mu_2 + \epsilon_i) \\ &= \beta_0' + \beta_1 X_{i,1} + \epsilon_i' \end{aligned}$$

Of course there could be more than one omitted variable. They would all get swallowed by the intercept and error term, the garbage bins of regression analysis.

$$
\begin{aligned}
Cov(X_{i,1}, \epsilon_i') &= Cov\left(X_{i,1}, (\beta_2 X_{i,2} - \beta_2 \mu_2 + \epsilon_i)\right) \\
&= \beta_2 \phi_{12}
\end{aligned}
$$

So if there are omitted independent variables that are related to both the dependent variable and the independent variables in the model, the covariance between independent variables and the error term is non-zero.

$$
\begin{aligned}
Cov(X_{i,1}, Y_i) &= Cov\left(X_{i,1}, (\beta_0' + \beta_1 X_{i,1} + \epsilon_i')\right) \\
&= \beta_1 Var(X_{i,1}) + Cov(X_{i,1}, \epsilon_i') \\
&= \beta_1 \phi_{11} + \beta_2 \phi_{12}
\end{aligned}
$$

Try to estimate β_1 using the mis-specified model

$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_{i,1}-\overline{X}_1)(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_{i,1}-\overline{X}_1)^2} \\
&= \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x^2} \\
&\xrightarrow{a.s.} \frac{\sigma_{xy}}{\sigma_x^2} \\
&= \frac{\beta_1\phi_{11}+\beta_2\phi_{12}}{\phi_{11}} \\
&= \beta_1+\beta_2\frac{\phi_{12}}{\phi_{11}}
\end{aligned}
$$

Moral of the story

- In an observational study, if there are omitted variables that are related to both the DV and IVs in the model, the results of a regression analysis have no necessary relation to reality.
- This is *almost always* the case.
- Still okay for pure prediction.
- In an experimental study, variables whose values are randomly assigned are unrelated to the omitted variables, and things are much better.

What if you measure all possible independent variables and "control" for them?

Control: Hold x_2 constant

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\frac{\partial E(Y|\mathbf{x})}{\partial x_1} = \beta_1$$

What if you measure all possible independent variables and "control" for them?

- Okay, as long as you can measure them all without error
- Otherwise the situation is similar to omitted variables.
- If IVs are measured with error, related to DV and correlated with IVs of interest - disaster
- Solution: Take measurement error into account in the data collection *and* the statistical model.

Watch out for measurement error in the independent variables

- When we test for a relationship “controlling” for some set of variables, we are seeking it in the conditional distributions - conditional on the values of the variables for which we are controlling.
- If the control variables are measured with error, the conditional distributions given the observed variables need not be the same as the conditional distributions given the true variables .

Suppose you are testing the relationship of age to BMI, controlling for exercise and calorie intake.

- Questionnaire measures are known to be inaccurate. People mis-report, and *not* by a constant amount.
- And, age is related to both independent variables, especially exercise
- Can't see the control variables clearly to hold them constant
- So even if age is unrelated to BMI for every combination of *true* exercise and *true* calorie intake, a relationship can exist conditionally upon *observed* exercise and *observed* calorie intake.

Want to test B controlling for A: The poison combination

- A is related to the dependent variable
- A and B are related to each other, and
- A is measured with error

- Estimation of B's relationship with Y is biased
- Type I error can be badly inflated (Brunner and Austin, 2009)
- Large sample size makes it *worse*!
- For observational studies, all three conditions usually are present.

Especially a problem in observational medical research

- Seek to assess potential risk factors, controlling for known risk factors
- The known risk factors do matter
- Known and potential risk factors are correlated
- Known risk factors are difficult to measure without error
- Experimental research is essential to confirm findings - and it often does not.

But all is not lost

- As long as you are interested in prediction rather than interpretation, there is no problem. Test for whether age is a useful predictor is still valid, even if its usefulness comes from its correlation with true exercise.
- The problem comes from trying to use regression as a *causal* model for observational data.
- If one or more categorical independent variables are experimentally manipulated, analysis of covariance can help reduce MSE and makes the analysis more precise, even if the covariates (control variables) are measured with error.
- No inflation of Type I error rate for ANCOVA - because random assignment breaks up the association between A and B.

If it's an observational study, just ask

- How did you control for ____?
- How did you take measurement error into account? (There are ways, but if it were easy people would do it more often. Nature of data collection is involved, not just statistical analysis.)

- If they say "Oh, there was just a little measurement error," observe that if the sample is large enough, no amount of measurement error is safe. Brunner and Austin (2009) give a proof.
- If they say "Well, it's the best we could do," you could ask whether it's better to say something incorrect, or to be silent.

In this course

- We will carry out classical regression analysis on observational data *only* when our primary purpose is prediction.
- We will be very careful about the way we describe the results.
- We will use regression methods extensively on experimental data.