Regression of the Multiple Linear kind
Applied Stats Algorithm

Scientific question?

Classify Study

Response Variable

Multi-Var

Univariate

Numerical

Censored

Complete

Predictor(s)

Today

Categorical

Both

Categorical

Fixed Effects

Random Effects

2+ Factors
Regression (Historically)

- 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.”
Relationship of Variables

- **Y**: *dependent / response* variable
  - variable whose behavior we want to study
- **X**: *independent / explanatory / predictor* variable
  - variable used to help us study Y

**Functional relationship**
- $Y = f(X)$, where $f()$ is some function

**Statistical Relationship**
- $Y = f(X) + \varepsilon$, where $\varepsilon$ is statistical *error* term
Relationship of Variables

- Scatter plots of data pairs \((Y_i, X_i)\)

Functional Relationship

Statistical Relationship

Statistical Error
Regression Models

- Regression model describes the statistical relationship between the response variable $Y$, and one (or more) explanatory variables
  - The response variable $Y$ has a tendency to change with the explanatory variable $X$
  - The data are scattered around the regression curve
- Regression model assumes a distribution for $Y$ at each level of $X$
Uses of Regression Models

Regression models have three general uses:

- **Description**: How does salary change with years of service?
- **Control**: Is a particular salary too big, with relation to the years of service?
- **Prediction**: What will your salary be after 3 years of service?
The means of the distributions of $Y$ at different levels of $X$ follow the regression curve.
Regression Models

- “It makes sense to base inferences or conclusions only on valid models”
  - S. Sheather

- “All models are wrong, but some are useful”
  - G. Box
Simple Linear Regression

- Model for the (statistical) relationship between $Y$ and a *single* $X$
- The regression curve is a *straight* line

\[ f(X) = \beta_0 + \beta_1 X \]
Simple Linear Regression

- Formal model description:

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

- \( Y_i \) is the response value (random variable)
- \( X_i \) is the predictor value (known and constant)
- \( \beta_0 \) is the Y-intercept (constant parameter)
- \( \beta_1 \) is the slope (constant parameter)
- \( \epsilon_i \) is the error term (random variable)
Simple Linear Regression

- Assumptions:
  - $\varepsilon_i$ is RV with $E\{\varepsilon_i\} = 0$ and $V\{\varepsilon_i\} = \sigma^2$
  - for $i \neq j$, $\varepsilon_i$ and $\varepsilon_j$ are uncorrelated: $Cov\{\varepsilon_i, \varepsilon_j\} = 0$

- Implications for $Y$
  - $E\{Y_i\} = \beta_0 + \beta_1X_i \Rightarrow E\{Y\} = f(X) = \beta_0 + \beta_1X$
  - $V\{Y_i\} = \sigma^2$
  - $Cov\{Y_i, Y_j\} = 0$
Multiple Linear Regression

● Multiple regression in matrix form

\[ Y_i = \beta_0 + \beta_1 \cdot X_{i,1} + \cdots + \beta_{p-1} \cdot X_{i,p-1} + \varepsilon_i, \quad i = 1, \ldots, n \]

\[ \iff Y_{n\times1} = X_{n\times p} \cdot \beta_{p\times1} + \varepsilon_{n\times1} \]

● Where:

\[ \varepsilon_{n\times1} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad E\{\varepsilon_{n\times1}\} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = 0_{n\times1}, \quad V\{\varepsilon_{n\times1}\} = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I_{n\times n} \]
Multiple Linear Regression

- **Design matrix**

\[
X_{n \times p} = \begin{bmatrix}
1 & X_{1,1} & \cdots & X_{1,p-1} \\
\vdots & \vdots & \ddots & \vdots \\
1 & X_{n,1} & \cdots & X_{n,p-1}
\end{bmatrix} = \begin{bmatrix}
x_0 \\
x_1 \\
\vdots \\
x_{p-1}
\end{bmatrix}
\]

- **Parameters** \( \beta_{p \times 1} = \begin{bmatrix} \beta_0 & \cdots & \beta_{p-1} \end{bmatrix}' \)

- **Observations**

\[
Y_{n \times 1} = \begin{bmatrix} Y_1 \\
\vdots \\
Y_n \end{bmatrix}, \quad E \left\{ Y_{n \times 1} \right\} = X\beta, \quad V \left\{ Y_{n \times 1} \right\} = \sigma^2 I_{n \times n}
\]
Least Squares Estimation

- **Parameter Estimates**
  \[ b_{p \times 1} = \left( X'X \right)^{-1} X'Y \]

- **Fitted Values**
  \[ \hat{Y}_{n \times 1} = Xb = \left[ X X' \right]^{-1} X'Y = HY \]

- **Residuals**
  \[ e_{n \times 1} = Y - \hat{Y} = Y - HY = \left( I - H \right)Y \]
Distributional Results

- Parameter Estimates (normal errors)

\[ \mathbf{b}_{p \times 1} \sim N \left( \mathbf{\beta}, \sigma^2 \left( \mathbf{X}'\mathbf{X} \right)^{-1} \right) \]

- Fitted Values

\[ \hat{\mathbf{Y}}_{n \times 1} \sim N \left( \mathbf{X}\mathbf{\beta}, \sigma^2 \mathbf{H} \right) \]

- Residuals

\[ \mathbf{e}_{n \times 1} \sim N \left( \mathbf{0}, \sigma^2 \left( \mathbf{I} - \mathbf{H} \right) \right) \]
Analysis of Variance

\[ SSTO = \sum (Y_i - \bar{Y})^2 = Y' \left( I - \frac{1}{n} J \right) Y, \quad df = n - 1 \]

\[ SSE = \sum (Y_i - \hat{Y}_i)^2 = Y' (I - H) Y, \quad df = n - p \]

\[ SSR = \sum (\hat{Y}_i - \bar{Y})^2 = Y' \left( H - \frac{1}{n} J \right) Y, \quad df = p - 1 \]

where \( J_{n \times n} = [1]_{i,j=1,\ldots,n} \)
Diagnostics for Residuals

- Use residuals to study the following departures from normal linear regression
  1. Regression function is not linear
  2. Error terms are not independent
  3. Error terms do not have constant $\sigma$
  4. Presence of outliers (extreme departures from model)
  5. Error terms are not normally distributed
  6. One or more important predictor variables are omitted from model
Checking for Linearity

- To check for nonlinearity:
  - Plot residuals vs predictor variable ($e_i$ vs $X_i$)
  - Plot residuals vs fitted values ($e_i$ vs $\hat{Y}_i$)

- You can also just use scatter plot ($X_i$ vs $Y_i$), but residual plots are preferable because
  - Can spot nonlinearity more easily
  - Can check other assumptions
  - Applies to multiple linear regression
    - ($e_i$ vs $X_i$) is equivalent to ($e_i$ vs $\hat{Y}_i$) in simple regression, but not in multiple regression
Residual Plot

Linear Relationship

Scatter plot

Residual plot

Fitted line
Residual Plot

Non-linear Relationship

Scatter plot

Residual plot

Actual curve

Fitted line
Checking for Error Independence

- Residual sequence plot (vs time or distance)
  - Check for temporal or spatial dependence
Checking for Constant Variance

- Residual or absolute residual plot
  - Should be spread out evenly across $X$

![Residual Plot](image1)

![Absolute Residual Plot](image2)
Checking for Outliers

- Semi-studentized residuals:
  - Regression errors ($\varepsilon$) have variance $\sigma^2$
  - Residuals are estimates of errors, and MSE is estimate of $\sigma^2$
  - Semi-studentized residuals are “semi-standardized” residuals
    
    $$ e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}} $$

  - $e_i^*$’s variance should be $\approx 1$
Checking for Outliers

- Studentized residuals:
  - Regression errors (ε) have variance σ²
  - Residuals actually have variance σ²(1-h_{ii})
  - Studentized residuals are properly standardized
    
    \[ t_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}} \]
    
    - This is actually an *internally studentized* residual
    - If the i^{th} case is deleted (changes estimate of σ²), called *externally studentized*
Checking for Outliers

- Residual or studentized residual plot
  - Rule of thumb: observation $i$ is an outlier if $|e_i^*| > k$ where $k$ is inconsistently $\epsilon(2,4)$
Checking for Error Normality

- Look at 2 possible deviations from Normality
  - Assymetry:
  - Heavy tails:

- Not critical for fitting regression line, but important for some applications, especially predictions
Checking for Error Normality

- Symmetry: Any plot describing the *shape* of $e_i$'s distribution (box-plot, histogram, dot-plot,..)
Checking for Error Normality

- Heavy Tails: Normal QQ-plot

Right Tail  
Left Tail  
Both Tails

Normal Quantiles
Checking for Important Predictor Variables

- Residual plot vs possible predictor variables
  - Fit $Y = b_0 + b_1X + e$, and plot $e$ vs variable $Z$

Scatter plot

Residual plot on $Z$
To R Studio ...
Remedial Measures

- Even if one or more of the diagnostics show problems, the fitted model may be improved.
  
  1. **Regression function is not linear**
     
     In some cases, a *variable transformation* can make the data “more linear”. If not, an altogether different (e.g. nonlinear) model might be better.

  2. **Error terms are not independent**
     
     If dependence can be modeled (e.g. serially correlated $\varepsilon$’s), it can be incorporated in regression.
Remedial Measures

3. Error terms do not have constant $\sigma$
   If you can model variance as a function of $X$, a weighted regression can be used. Sometimes, variable transformations also stabilize variance.

4. Presence of outliers
   Sometimes outliers can be omitted from the data. Alternatively, different fitting methods can be used to reduce their impact (robust regression)
Remedial Measures

5. Error terms are not normally distributed
   Often, variable transformations can help. If not, try modelling the error term with a different distribution

6. One or more important predictor variables are omitted from model
   Include them in the model (multiple regression)

   - And … there are always randomization tests
Linear Regression Pitfalls

Simpson’s Paradox
Omitted Variables
Regression to the Mean
Multiple Linear Regression

- The standard elementary tests all have a single independent variable, so they should be used with caution in practice

- Example
  - Artificial and extreme, to make a point

- Suppose the correlation between Age and Strength is $r = -0.96$
This is an example of Simpson’s paradox.

The overall relationship between variables is clear, but it is reversed when examined separately for the values of another variable.

In the example, species is a confounding variable (2 criteria).

It happens with real data.

More to come in the unit on Categorical Data Analysis.
Simpson’s Paradox

- Can be hard to see when there are lots of variables
- Need a systematic way to allow (control) for potential confounding variables by including them in the analysis
Three Meanings of Control

- Experimental
- Sub-division
- Model-based

Multiple regression is the prime example of model-based Control.
Least Squares Line
Least Squares Plane

\[ \hat{Y} = b_0 + b_1 x_1 + b_2 x_2 \]
What is $b_2$?

\[ \hat{Y} = b_0 + b_1 x_1 + b_2 x_2 \]

- Hold $x_1$ constant at some fixed level
- What is predicted $Y$, as a function of $x_2$?

\[ \hat{Y} = (b_0 + b_1 x_1) + b_2 x_2 \]

- $b_1 x_1$ is now part of the intercept, $b_2$ the slope
- It’s the rate at which predicted $Y$ changes as a function of $x_2$, with $x_1$ held constant
- Say “controlling” for $x_1$
Control for $x_1$, $x_3$ and $x_4$

\[
\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4
\]

\[
\hat{Y} = (b_0 + b_1 x_1 + b_3 x_3 + b_4 x_4) + b_2 x_2
\]
Significance test for $b_k$

- Test for $b_k$ tells you whether $x_k$ makes a meaningful contribution to predicting $Y$, controlling for the other independent variables

- “Allowing for”
- “Holding constant”
- “Blocking out”
Statistical Model

- There are $p - 1$ independent variables
- For each combination of IVs, the conditional distribution of the dependent variable $Y$ is Normal, with constant variance
- The conditional population mean of $Y$ depends on the $X$ values, as follows:

$$E[Y | X = x] = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$
Control means hold constant

\[ E[Y|X = x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \]

\[ \frac{\partial}{\partial x_3} E[Y|X = x] = \beta_3 \]

- So \( \beta_3 \) is the rate at which \( E[Y|X] \) changes as a function of \( x_3 \) with all other variables held constant at fixed levels.
Unit change in $x_3$, holding other variables constant

- The difference is:
  $$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_3 + 1) + \beta_4 x_4 \nonumber$$
  $$- \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \nonumber$$
  $$= \beta_3 \nonumber$$

- So $\beta_3$ is the amount that $E[Y|X]$ changes when $x_3$ is increased by one unit, and all other variables held constant at fixed levels.
Conditional Distributions are Normal

\[ E[Y|X = x] = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1} \]

- Same variance, and population mean
- This means the only way \( Y \) can be related to any \( x \) is through the \( \beta \) values.
- Because the (conditional) expected value has a simple structure, it is possible to draw conclusions about the conditional distribution of \( Y \), holding the IVs constant at sets of \( x \) values where there are no data!
High School Calculus and University Calculus

\[ \hat{Y} = -84.85 + 1.79x \]

- Sub-division approach needs lots of data at a particular value to give a good estimate of the conditional population mean.
- Here, we can easily give a good estimate of university calculus mark for a HS Calculus mark of 59, (estimate is 20.76) even though there was just one person with a 59 in the data and he dropped the course.
- We can do this because of the assumption (model) \( E(Y|x) = \beta_0 + \beta_1x \)
- The more data you have, the less you need to assume.
Statistics b estimate parameters $\beta$

$$E[Y|X = x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

- All of the same comments apply to statistics as well as the parameters
  - ie. $b_1$ is the estimated change in $E[Y]$ for a unit change in $x_1$, holding everything constant
Analysis of Variance

- Variation to explain: **Total Sum of Squares**
  \[ SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \]

- Variation that is still unexplained: **Error Sum of Squares**
  \[ SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \]

- Variation that is explained: **Regression (or Model) Sum of Squares**
  \[ SSR = SSTO - SSE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \]
# ANOVA Summary Table

## Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$p - 1$</td>
<td>$SSR$</td>
<td>$MSR = SSR/(p - 1)$</td>
<td>$F = \frac{MSR}{MSE}$</td>
<td>$p$-value</td>
</tr>
<tr>
<td>Error</td>
<td>$n - p$</td>
<td>$SSE$</td>
<td>$MSE = SSE/(n - p)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>$SSTO$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

\[ R^2 = \frac{SSR}{SSTO} \]
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

\[ R^2_F \geq R^2_R \]
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 $s$ of them) equal zero

$$F = \frac{(SSR_F - SSR_R)/s}{MSE_F}$$
Coefficient of Partial Determination

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

$$R^2_{Partial} = \frac{R^2_F - R^2_R}{1 - R^2_R} = \alpha$$

- Increase in explained variation expressed as a fraction of the variation that the reduced model does not explain.
Coefficient of Partial Determination

Full model: \( E\{Y\} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 \)

Reduced model: \( E\{Y\} = \beta_0 + \beta_1X_1 + \beta_2X_2 \)

- Measure relative reduction in Y variance after introducing \( X_3 \) to model with \( X_1, X_2 \)

\[
R^2_{Y3|1,2} = \frac{SSR(X_3 \mid X_1, X_2)}{SSE(X_1, X_2)} = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{SSE(X_1, X_2)}
\]

- Takes values in \([0,1]\)
- \( R^2 \) of regressing residuals of reduced model to residuals of \( E\{X_3\} = \beta_0 + \beta_1X_1 + \beta_2X_2 \)
For any given sample size, the bigger $a$ is, the bigger $F$ becomes.

For any $a \neq 0$, $F$ 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.

$$F = \left( \frac{n - p}{s} \right) \left( \frac{a}{1 - a} \right)$$
Can express $a$ in terms of $F$

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

- Often, scientific journals just report $F$, numerator df = $s$, 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 subject to rounding error than the one in terms of R-squared values.
When you add independent variables to a model

- 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 consistent with what happened in the age and strength example.
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 \( \epsilon_i \) independent of \( X_{i,1} \) and \( X_{i,2} \).

- \( X_2 \) is not observed
Omitted Variables

Since $X_2$ is not observed, it is absorbed

\[ Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \]
\[ = (\beta_0 + \beta_2 \mu_2) + \beta_1 X_{i,1} + (\beta_2 X_{i,2} - \beta_2 \mu_2 + \epsilon_i) \]
\[ = \beta'_0 + \beta_1 X_{i,1} + \epsilon'_i \]

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

\[
\text{Cov}(X_{i,1}, e_i') = \text{Cov}(X_{i,1}, (\beta_2 X_{i,2} - \beta_2 \mu_2 + \epsilon_i)) = \beta_2 \phi_{12}
\]

- So if there are omitted IVs that are related to both the response variable and the IVs in the model, the covariance between IVs and the error term is non-zero.
Omitted Variables

\[ \text{Cov}(X_{i,1}, Y_i) = \text{Cov}(X_{i,1}, (\beta'_0 + \beta_1 X_{i,1} + \epsilon'_i)) \]
\[ = \beta_1 \text{Var}(X_{i,1}) + \text{Cov}(X_{i,1}, \epsilon'_i) \]
\[ = \beta_1 \phi_{11} + \beta_2 \phi_{12} \]
Omitted Variables

- Try to estimate $\beta_1$ using the mis-specified model

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_{i,1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_{i,1} - \bar{X}_1)^2}
\]

\[
= \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}
\]

\[
a.s. \rightarrow \frac{\sigma_{xy}}{\sigma_x^2}
\]

\[
= \frac{\beta_1 \phi_{11} + \beta_2 \phi_{12}}{\phi_{11}}
\]

\[
= \beta_1 + \beta_2 \frac{\phi_{12}}{\phi_{11}}
\]
In an observational study

- If there are omitted variables that are related to both the response variable and IVs in the model, the results of a regression analysis have no necessary relation to reality.

This is almost always the case!

Still OK for pure prediction

In an experimental study

- Variables whose values are randomly assigned are unrelated to the omitted variables, and things are much better.
Control

- Hold $X_2$ constant

\[ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \]

\[ E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \]

\[ \frac{\partial E(Y|x)}{\partial x_1} = \beta_1 \]
What if you measure all possible IVs and “control” for them

- OK, 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
  - Structural Equation Modelling
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.
Example

- Suppose you are testing the relationship of age (Y) to BMI (X), controlling for exercise and calorie intake.
- Questionnaire measures are known to be inaccurate. People mis-report, and not by a constant amount.
  - Measurement error
- Age is related to both covariates, especially exercise
Example

- 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 $X_1$ controlling for $X_2$: The poison combination

- $X_2$ is related to the dependent variable
- $X_1$ and $X_2$ are related to each other, and
- $X_2$ is measured with error

- Estimation of $X_1$’s relationship with $Y$ is biased
- Type I error is badly inflated
  - Brunner and Austin, 2009
- Large sample size makes it worse!
- For observational studies, all three conditions are usually 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.
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.
ANCOVA

- 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.
Regression to the mean

\[ Y = X\beta + \varepsilon \]

\[ \hat{\beta} = (X'X)^{-1}X'Y \]

\[ \hat{Y} = X\hat{\beta} \]

\[ e = Y - \hat{Y} \]
Measure the same thing twice, with error

\[ Y_1 = X + \epsilon_1 \]
\[ Y_2 = X + \epsilon_2 \]

\[ X \sim N(\mu, \sigma_x^2) \]
\[ \epsilon_1 \perp \epsilon_2 \sim N(0, \sigma_e^2) \]
\[ X \perp \epsilon \]

\[
\begin{pmatrix}
Y_1 \\
Y_2
\end{pmatrix}
\sim
N
\left( 
\begin{bmatrix}
\mu \\
\mu
\end{bmatrix},
\begin{bmatrix}
\sigma_x^2 + \sigma_e^2 & \sigma_x^2 \\
\sigma_x^2 & \sigma_x^2 + \sigma_e^2
\end{bmatrix}
\right)
\]
Conditional distribution of $Y_2$ given $Y_1 = y_1$

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

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

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

where $\rho = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}$. 
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 is exactly the “regression toward the mean” that Galton observed.
Regression toward the mean

- Does not imply systematic change over time
- Is just 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 the worst on some 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 the first time – completely artificial!
A simulation study

- Measure something twice with error – 500 observations
- Select the best and worst 50 observations
- Do two-sided matched t-tests at $\alpha = 5\%$
  - What proportion of the time do the worst 50 show significant average improvement?
  - What proportion of the time do the best 50 show significant average deterioration?
To R Studio ...
Summary

- 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 differences from baseline may still be useful