PSC 504: Regression

Matthew Blackwell

2/28/2013

Regression is probably the most commonly used statistical tool in the social sciences. It's usually the first technique you learn because it has nice properties and simple interpretations. Regression is one way to estimate causal effects, but it comes with a good deal of baggage. Most methodological developments in the social sciences have been reactions to linear regression models estimated by ordinary least squares (OLS). Thus, we often come to regression with the feeling that its assumptions are too strong and that we should model the full distribution of the data. And this can be a good approach! But the simple truth is that regression has many good properties even when those burdensome assumptions do not hold (especially in large samples). Freedman and many other statisticians have warned us that regression is not justified by randomization, but others (Angrist and Pischke 2008) believe that if a causal effect is identified, then a regression is not a bad way to approximately estimate that effect. Of course, the is as it always is: everyone is right and it depends. Today we'll try to salvage regression from the ashes of 1980's textbooks and see how it handles or mishandles causal inference.

Agnostic views on regression

- When we usually learn regression, we state a series of assumptions such as linearity, fixed regressors, Normal errors, homoskedastic error variance, and so on. We then use these assumptions to show that the ordinary least squares estimator for a linear model is optimal in a certain sense. That is, we could write down a model for our data:

$$[Y_i|X_i] \sim \mathcal{N}(X_i\beta, \sigma^2)$$

- This model assumes that the distribution of Y is normal with a mean that is a linear function of the covariates, with a common variance, σ^2, across units. In this siutation, we can show that the maximum likelihood estimator for β is the usual OLS estimator, $\hat{\beta}_{\text{OLS}}$:

$$\hat{\beta}_{\text{OLS}} = \left[\sum_{i=1}^{N} X_i X_i'\right]^{-1} \sum_{i=1}^{N} X_i Y_i$$

- Not only is $\hat{\beta}_{\text{OLS}}$ the MLE for this model, it is also the best linear unbiased estimator. These are great properties, but they depend heavily on the distributional assumptions that we have made. If some of these assumptions fail to hold, we lose these properties.

- Instead of taking these assumptions as gospel and interpretting our regressions as such, we can alternatively take an "agnostic" view of regression. We need not take the assumptions of the linear model seriously in order to use the regression machinery to estimate causal effects.

Conditional Expectation Function

- Instead of formulating a model for the distribution of Y_i, we will attempt to estimate how the distribution of Y_i changes with a covariate X_i. That is, we will use X_i to predict the value of Y_i. A useful tool for this prediction is the conditional expectation function (CEF), which is the mean (or population average) of Y_i when X_i is held fixed. Obviously, the CEF is a function of X_i and we write it as $E[Y_i|X_i]$ and we can derive it from the conditional distribution of Y_i:

$$E[Y_i|X_i = x] = \sum_t t \Pr[Y_i = t|X_i = x]$$

- Remember that we can use the law of iterated expectations to recover the unconditional expectation of Y_i: $E[E[Y_i|X_i]]$.

- We can decompose Y_i into the CEF and a residual that is mean independent of X_i.

$$Y_i = E[Y_i|X_i] + e_i \quad E[e_i|X_i] = 0$$

- The CEF is the function of X_i that best predicts (in a mean squared error sense) Y_i.

- The Law of total variation:

$$V(Y_i) = E[V(Y_i|X_i)] + V(E[Y_i|X_i])$$

Justifying linear regression

- Define linear regression:

$$\beta = \arg\min_b E[(Y_i - X_i'b)^2]$$

- The solution to this is the following:

$$\beta = E[X_i X_i']^{-1} E[X_i Y_i]$$

- We can define the residuals $e_i = Y_i - X_i'\beta$, which are uncorrelated with the regressors. Again, no assumptions here yet.

- A little regression anatomy that will be useful later when we focus on the coefficient for the treatment. First, let $\tilde{X}_{ki}$ be the residual from a regression of X_{ki} on all the other variables:

$$\beta_k = \frac{\text{Cov}(Y_i, \tilde{X}_{ki})}{V(\tilde{X}_{ki})}$$

- Justification 1: if the CEF is linear, the population regression function is it. That is, if $E[Y_i|X_i] = X_i'b$, then $b = \beta$.

- When would we expect the CEF to be linear? Two cases. One is if the data (the outcome and covariates) are multivariate Normal. The other is if the linear regression is **saturated**. A saturated regression model is one in which there is a parameter for each unique combination of the covariates. In this case, the regression fits the CEF perfectly because the CEF is a linear function of the dummy categories.

- Imagine a regression model with two binary variables, X_{1i} and X_{2i}. One could represent incumbent status of a candidate and the the other could be the party of that candidate. Then, a saturated regression model would be one that has each of these variables and their interaction. In this case, there are four different values that the CEF can take:

$$E[Y_i|X_{1i} = 0, X_{2i} = 0]$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 0]$$
$$E[Y_i|X_{1i} = 0, X_{2i} = 1]$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 1]$$

- We can write the CEF as follows:

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- There are no restrictions on this CEF. We can see this by noting that there are as many parameters as combinations of the covariates:

$$E[Y_i|X_{1i} = 0, X_{2i} = 0] = \alpha$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 0] = \alpha + \beta$$
$$E[Y_i|X_{1i} = 0, X_{2i} = 1] = \alpha + \gamma$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 1] = \alpha + \beta + \gamma + \delta$$

- Thus, a linear regression with the same set of covariates will perfectly fit the CEF because each value of the CEF is being estimated independently. A saturated model like this is the most flexible linear regression model we can have. Dropping interaction terms or making linearity assumptions on multi-category variables will cause the regression model to be no longer saturated.

- Also note that a saturated model perfectly fits the CEF without making any assumptions on the distribution of Y_i. Linearity of the CEF is not an assumption here, it's a fact about saturated CEFs. Thus, a saturated linear regression with a binary or non-negative outcome is perfectly valid: this regression will accurately recover the parameters of the CEF for these variables.

- Justification 2: $X_i'\beta$ is the best linear predictor (in a mean-squared error sense) of Y_i.

- Justification 3: $X_i'\beta$ provides the minimum mean squared error linear approxmiation to $E[Y_i|X_i]$.

- Thus, even if the CEF is not linear, a linear regression provides the best linear approximation to that CEF. This is crucial to an agonostic view of regression: we don't need to believe the assumptions (linearity) in order to use regression as a good approximation to the CEF.

Asymptotic OLS inference

- $Y_i = X_i'\beta + [Y_i - X_i'\beta] = X_i'\beta + e_i$. Note the residual e_i is uncorrelated with X_i:

$$
\begin{aligned}
E[X_i e_i] &= E[X_i(Y_i - X_i'\beta)] \\
&= E[X_i Y_i] - E[X_i X_i'\beta] \\
&= E[X_i Y_i] - E[X_i X_i' E[X_i X_i']^{-1} E[X_i Y_i]] \\
&= E[X_i Y_i] - E[X_i X_i'] E[X_i X_i']^{-1} E[X_i Y_i] \\
&= E[X_i Y_i] - E[X_i Y_i] = 0
\end{aligned}
$$

- We get this without making any assumptions on the linearity of $E[Y_i|X_i]$. We can still think of $X_i'\beta$ as the best linear approximation to $E[Y_i|X_i]$, no matter what the CEF looks like (that is, linear or non-linear).
- With this representation in hand, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i e_i$$

- Using the standard tools of asymptopia (Slutsky's theorem, central limit theorem), we know that $\sqrt{N}(\hat{\beta} - \beta)$ coverges in distribution to a Normal distribution with mean vector 0 and covariance matrix, Ω:

$$\Omega = E[X_i X_i']^{-1} E[X_i X_i' e_i^2] E[X_i X_i']^{-1}.$$

- We can replace e_i with its emprical counterpart $\hat{e}_i = Y_i - X_i'\hat{\beta}$ and replace the population moments of X_i with their sample counterparts to estimate this covariance matrix. The square root of the diagonals of this covariance matrix are the "robust" or Huber-White standard errors that Stata commonly report.
- Note that heteroskedaticity will definitely occur when $E[Y_i|X_i]$ is not linear, but we use the linear regression to approxmiate it. It will also crop up if the CEF is linear, but the conditional variance is different for different values of X_i.
- Again, this is a large sample property of linear regression that does not rely on the linearity of the CEF. These results do rely on (Y_i, X_i) being a simple random sample from a population and obviously they are large sample properties.

Regression and causality

- When we look at a textbook, we often see regression defined without respect to causality. There is talk of the $\hat{\beta}$ estimator being "biased," but it isn't always clear what the "correct" specification would look like. There is an implicit assumption of causality, but no formal definitions. This can obscure the identification of the causal effects of interest. Today, we'll see if we can estimate causal effects with regression.
- Angrist and Pischke (2008) argue that a regression is causal when the CEF it approximates is causal. Identification is king.
- We will show that under certain conditions, a regression of the outcome on the treatment and the covariates can recover **a** causal parameter, but perhaps not the one in which we are interested.

- Quick reminder: we have potential outcomes, $Y_i(1)$ and $Y_i(0)$, and two parameters, the ATE and ATT:

$$\begin{aligned}\tau &= E[Y_i(1) - Y_i(0)],\\ \tau_{\text{ATT}} &= E[Y_i(1) - Y_i(0)|A_i = 1].\end{aligned}$$

- We have shown in past weeks that these effects are identified when ignorability holds. Angrist and Pischke (2008) call this the conditional independence assumption (CIA).

Linear constant effects model, binary treatment

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$\begin{aligned}Y_i &= A_iY_i(1) + (1 - A_i)Y_i(0)\\ &= Y_i(0) + (Y_i(1) - Y_i(0))A_i\\ &= \mu^0 + \tau A_i + (Y_i(0) - \mu^0)\\ &= \mu^0 + \tau A_i + v_i^0\end{aligned}$$

- Note that if ignorability holds (as in an experiment) for $Y_i(0)$, then it will also hold for v_i^0, since μ^0 is constant. Thus, this satifies the usual assumptions for regression.

- Let's now say that ignorability holds only conditional the covariates, so $Y_i(a) \perp\!\!\!\perp A_i|X_i$. We will assume a linear model for the potential outcomes:

$$Y_i(a) = \alpha + \tau a + \eta_i$$

- Because we are assuming the effect of A is constant here, the η_i are the only source of individual variation and we have $E[\eta_i] = 0$. We can use the consistency assumption to write this as a linear regression model:

$$Y_i = \alpha + \tau A_i + \eta_i.$$

- Now, we know that ignorability only holds on X_i. Let's assume that conditional expectation of η_i is linear in the covariates: $E[\eta_i|X_i] = X_i'\gamma$. With this, we can write $\eta_i = X_i'\gamma + \nu_i$, with $E[\nu_i|X_i] = 0$. This is an assumption: it maybe the case that the CEF of the error is not linear in the covariates. This is one of those strong modeling assumptions that we wanted to avoid when we were using matching. We'll look at non-parametric regression below that relax this assumption. For now, we can see that (using ignorability and consistency in the first equality):

$$\begin{aligned}E[Y_i|A_i, X_i] = E[Y_i(a)|X_i] &= \alpha + \tau A_i + E[\eta_i|X_i]\\ &= \alpha + \tau A_i + X_i'\gamma + E[\nu_i|X_i]\\ &= \alpha + \tau A_i + X_i'\gamma\end{aligned}$$

- Thus, a regression where A_i and X_i enter linearly will correctly estimate the average treatment effect, τ, since the residual of the linear regression is independent of the covariates:

$$Y_i = \alpha + \tau A_i + X_i'\gamma + \nu_i$$

- Note that nothing we have done changes if A_i were continuous or ordinal with multiple categories. As long as the potential outcomes follow a linear model with a constant effect, the above will hold.
- Let's the review the assumptions that allow "off-the-shelf" regression to identify the ATE: ignorability, constant effect of treatment, and linear separability in the covariates.

Heterogeneous effects, binary treatment

- Experiment:

$$\begin{aligned}
Y_i &= A_i Y_i(1) + (1 - A_i) Y_i(0) \\
&= Y_i(0) + (Y_i(1) - Y_i(0)) A_i \\
&= \mu^0 + (Y_i(1) - Y_i(0)) A_i + v_i^0 \\
&= \mu^0 + (Y_i(1) - Y_i(0)) A_i - v_i^1 A_i + v_i^1 A_i + v_i^0 \\
&= \mu^0 + (\mu^1 - Y_i(0)) A_i + v_i^1 A_i - v_i^0 A_i + v_i^0 A_i + v_i^0 \\
&= \mu^0 + (\mu^1 - \mu^0) A_i + v_i^0 + (v_i^1 - v_i^0) A_i \\
&= \mu^0 + \tau A_i + \varepsilon_i
\end{aligned}$$

Heterogeneous effects and matching estimators

- Let's relax the assumption of constant effects. First, it's useful to remember that the ATE is a weighted sum of the conditional ATEs: $\tau(x) = E[Y_i(1) - Y_i(0) | X_i = x]$. We compute the overall ATE as $\tau = \sum_x \tau(x) \Pr[X_i = x]$. Note that, by ignorability, $\tau(x) = E[Y_i(1) | X_i = x, A_i = 1] - E[Y_i(0) | X_i = x, A_i = 0]$.
- There is a similar derviation for the ATT:

$$\begin{aligned}
\tau_{ATT} &= E[Y_i(1) - Y_i(0) | A_i = 1] \\
&= E\left\{ E[Y_i(1) - Y_i(0) | X_i, A_i = 1] | A_i = 1 \right\} \\
&= E\left\{ E[Y_i(1) | X_i, A_i = 1] - E[Y_i(0) | X_i, A_i = 1] | A_i = 1 \right\} \\
&= E\left\{ E[Y_i(1) | X_i, A_i = 1] - E[Y_i(0) | X_i, A_i = 0] | A_i = 1 \right\} \\
&= E[\tau(x) | A_i = 1] \\
&= \sum_x \tau(x) \Pr[X_i = x | A_i = 1]
\end{aligned}$$

- Remember Bayes's formula: $\Pr[X_i = x | A_i = 1] \Pr[A_i = 1] = \Pr[A_i = 1 | X_i = x] \Pr[X_i = x]$, so we can rewrite the ATT as a propensity score-weighted function of the CATEs (with a normalizing factor):

$$\tau_{ATT} = \frac{\sum_x \tau(x) \cdot \Pr[A_i = 1 | X_i = x] \Pr[X_i = x]}{\sum_x \Pr[A_i = 1 | X_i = x] \Pr[X_i = x]}$$

Heterogeneous effects and regression

- Now that we can see that both the ATE and ATT are weighted averages of the CATEs. We will now show that the regression estimator with full saturation in X_i is also a weighted function of the CATEs.
- Suppose that we can write a dummy variable for each unique combination of X_i: $D_{xi} = \mathbb{I}(X_i = x)$, where x can be a vector. Thus, each combination of the covariates has its own binary variable. We can then define the following regression:

$$Y_i = \sum_x D_{xi}\alpha_x + \tau_R A_i + e_i.$$

- Again, this regression is saturated in the covariates, which means that it is linear in the covariates by construction. We are interested in seeing if τ_R is equal to the ATE or the ATT. The fact that τ_R does not vary with the units is not a constant effects assumption, it's just a parameter of this regression model.
- How can we investigate τ_R? Well, we can rely on the regression anatomy:

$$\tau_R = \frac{\text{Cov}(Y_i, \tilde{A}_i)}{V(\tilde{A}_i)}$$

- Remember that $\tilde{A}_i$ is the residual from a regression of A_i on the full set of dummies. Because that regression is saturated (because the dummies saturate the covariates), we know that these resdiuals are: $\tilde{A}_i = A_i - E[A_i|X_i]$.
- We also know that a regression of Y_i on the treatment and covariates is the same as a regression of the $E[Y_i|X_i, A_i]$ on the treatment and covariates. Thus, in the above expression, we can replace Y_i with $E[Y_i|X_i, A_i]$.

$$\tau_R = \frac{\text{Cov}(E[Y_i|X_i, A_i], A_i - E[A_i|X_i])}{V(A_i - E[A_i|X_i])} = \frac{E\left\{E[Y_i|X_i, A_i](A_i - E[A_i|X_i])\right\}}{E[(A_i - E[A_i|X_i])^2]}$$

- Why stop here? We can simplify the CEF a bit more:

$$\begin{aligned} E[Y_i|X_i, A_i] &= E[A_iY_i(1) + (1 - A_i)Y_i(0)|X_i, A_i] \\ &= E[Y_i(0)|X_i, A_i = 0] + A_iE[Y_i(1) - Y_i(0)|X_i, A_i] \\ &= E[Y_i|X_i, A_i = 0] + \tau(X_i)A_i \end{aligned}$$

- We can plug this into the numerator of τ_R above. After a bit of simplification, we get the following:

$$\tau_R = \frac{E\left[\tau(x)(A_i - E[A_i|X_i])^2\right]}{E[(A_i - E[A_i|X_i])^2]} = \frac{E[\tau(X_i)\sigma_A^2(X_i)]}{E[\sigma_A^2]}$$

- Here, σ_A^2 is the variance of A_i conditional on X_i. Note that, since A_i is binary, we know that:

$$\sigma_A^2 = \Pr[A_i = 1|X_i](1 - \Pr[A_i = 1|X_i])$$

- Thus, we can finally plug all of this in to find a representation of τ_R:

$$\tau_R = \frac{\sum_x \tau(x)\left[\Pr[A_i = 1|X_i = x](1 - \Pr[A_i = 1|X_i = x])\right]\Pr[X_i = x]}{\sum_x \left[\Pr[A_i = 1|X_i = x](1 - \Pr[A_i = 1|X_i = x])\right]\Pr[X_i = x]}$$

- Thus, under the assumption of saturated in the covariates, the coefficient on the treatment in a linear regression is a weighted average of the within-stratum effects. Remember that the ATT and ATE are also weighted average of the within-stratum effects, but they weight by the size of those strata (overall for the ATE or among the treated for the ATT). The within-stratum effects in regression are weighted by the conditional variance of treatment in that stratum.
- Why does the OLS estimator weight by the conditional variance of the treatment? OLS is a minimum-variance estimator. Thus, it gives more weight to strata with lower expected variance in their estimates. That is, it gives higher weight to more precise within-strata estimates. When are these estimates going to be more precise? When the treatment and control group are roughly the same size and so the variance is maximized. This is because $V(A|X = x) = \Pr[A|X = x](1 - \Pr[A|X = x])$. This is maximized with $\Pr[A|X = x] = \frac{1}{2}$.
- When does $\tau = \tau_R$? When $\tau(x) = \tau$ is constant across the strata of the covariates. In that case, we recover the results from before. The other case is if the probability of treatment is constant across the strata. In that case, we can pull out the $\Pr[A_i = 1|X_i = x]$ terms out of all the sums and they cancel, leaving us simply weighting by the size of each stratum.
- Example: two strata, one with $\tau(1) = 1$ and the other with $\tau(2) = -1$. We have 75% of people in stratum 1, so that the ATE is 0.5. We also have that the first group has 90% treated units, while the second stratum has 50% treated units. Then, the ATT is 0.685. The regression estimator will be 0.0384.
- Obviously this will be the biggest issue when
- OLS weight these weights intuitively gives weight zero to strata with only treated or control units, because those have 0 variance. But if the model isn't completely saturated in X_i, this result doesn't hold and some extraplation will occur.
- A version of this result holds when when generalized A_i to be ordinal or continuous. This case is more difficult because it involves average derivatives as opposed to simple differences.

Non-parametric regression

- What do we do about the fact that the regression coefficient does not estimate the ATE or the ATT under heterogeneous effects? Do we have to abandon regression? First, it's important to note that the bias of regression may be small in most situations, but it could be large, especially if we cannot saturate the model in X_i. Instead, we may want to turn to a different technique.
- An alternative regression estimator is sometimes called the imputation estimator and involves imputing the values of $Y_i(1)$ and $Y_i(0)$ for each unit, using a regression, and then taking the average of the differences between these imputations as the estimator for the ATE. Let $\hat{\mu}_a(x)$ be a consistent estimator for $\mu_a(x) = E[Y_i(a)|X = x]$. We could always run a saturated (in X_i) linear regression in the treated and control groups separately as this estimator.

$$\tau_{imp} = \frac{1}{N} \sum_i \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

- Thus, we use the regression(s) to predict values of the potential outcomes, then average across the imputed individuals treatment effects. Because each of the regression estimators are consistent, then the imputations estimator is consistent for the ATE as well.
- Why don't people use this more? Well, it isn't as easy to calculate as the dead-easy regression coefficients. Furthermore, the standard errors are not straightforward and usually require bootstrapping. Plus, under constant effects, this estimator is identical to the regression estimator.
- In general, we can run separate regressions: $\mu_a(x) = X_i'\beta_a$, which is identical to running regression with a full set of interaction between the treamtent and the covariates.
- The recent trend in the literature is to estimate $\mu_a(x)$ using non-parametric methods such as kernel regression. See Imbens (2004) for more about this.

Limited dependent variables

- Traditionally, we shy away from regression with limited dependent variables. We look to logits, probit, Tobits, MLE, etc, etc. We might wonder if this is necessary? Economists often just use OLS. Why the difference in opinion?

Experiments

- Note that, with a binary randomized treatment, we can just do a difference in means (aka regression) without imposing any assumptions on the distribution of Y_i. This is because a model like that is saturated.
- Running a probit model in this case requires us to make a non-linear transformation of the model parameters to get back to effects on the same scale as the dependent variable. The validity of this transformation depends on the model we choose and may depend crucially on constant effects assumptions.
- Even if there are covariates, if they are binary, we can estimate the CATE, $\tau(x)$ and then average those over the distribution of the covariates. Because each within-strata estimate is linear, the weighted average is also linear.
- Conditioning on the outcome being positive (COP) introduces post-treatment bias.
- There being a cutoff of zero, say in payments, this might substantively interesting, not a censoring problem.

Covariates and nonlinearity

- When we add covariates to a regression, linearity becomes an assumption. Angrist and Pischke (2008) give evidence that the difference between marginal effects calculated from non-linear models and OLS coefficients will roughly be the same, so we shouldn't worry too much about this.
- This isn't necessarily true, though, and things can go wrong in small samples or the covariates have a strange distribution. In this case, we might want to use the usual non-linear models. But it's important to calculate causal quantities such as the first difference or the average derivative, since these are on the

scale of the dependent variable. The coefficients on latent variables are non-causal and, in general, not that interesting.

- Use censoring models when there is true censoring, such as top-coding of income, not just a substantive zeroes.