Linear Regression Diagnostics and Inference Notes

Questions

- Somewhat philosophically, how do we know whether some regressor is included in the causal model or merely used as a proxy? For example, in a regression of hourly wage on degree status and age, is age included in the causal model or merely a proxy for work experience? Or is age a proxy, in a more extended way, for relevant work experience, or productivity? Similarly, should we think that degree status is a proxy for ability or a direct determinant?
  - Depends on the case, but it can also generally be argued both ways. There is no obviously correct answer. So, for example if degree and age are explicit requirements or we have reason to suspect explicit discrimination, these are causal. If holding knowledge and experience constant eliminates any variation or effect of degree and age, then degree and age are better treated as proxies.

Regression Diagnostics

- In these notes, the primary relationship of interest is the relationship between the population linear regression and the sample linear regression. The notation is revised from earlier notes such that we here denote the parameters and residual of a population linear regression using β and u rather than ρ and e as before. β and u in these notes do not refer to the parameters and residual of a causal model.
- The inclusion of an additional regressor that is independently correlated with the dependent variable reduces the magnitude of residuals, hence reduces the variance of residuals, and the increases the precision of the estimates of the other regression coefficients.
- A linear regression model that is overfit on data is one that is too complex or flexible in relation to the complexity and volume of data. A model that is overfit is excessively sensitive to noise in the data. Such a model will yield a low standard error of regression and high R^2 measure of fit, but will yield less accurate predictions for data outside the training set. For example, including higher-order polynomial regressors than necessary results in a model in which the higher-order terms dominate for large values of the dependent variable, hence a model that yields poor predictions for such large values.

Standard Error of Regression

- The standard error of regression $s_{\hat{u}}$ is an estimator of the population standard deviation of the residual u.
  - Recall that a population linear regression is some $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + u$, where $Y, X_1, X_2, \ldots$ corresponds to some arbitrary observation in the population, and $\mathbb{E}u = \mathbb{E}X_1 u = \mathbb{E}X_2 u = \ldots = 0$.
  - The corresponding sample linear regression, given some sample drawn from the population, is some $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{u}_i$, where $Y, X_{1i}, X_{2i}, \ldots$ corresponds to some arbitrary observation in the sample, and $\hat{\mathbb{E}}\hat{u} = \hat{\mathbb{E}}X_1\hat{u} = \hat{\mathbb{E}}X_2\hat{u} = \ldots = 0$.
  - The variance of the error term u in the population linear regression is $\sigma_u^2 := var(u)$. This is consistently estimated by the sample variance of the residual $\hat{u}$ in the sample linear regression, which is $s_{\hat{u}}^2 := \frac{n}{n-k-1}\hat{\mathbb{E}}(\hat{u} - \hat{\mathbb{E}}\hat{u})^2 = \frac{n}{n-k-1}\hat{\mathbb{E}}\hat{u}^2$, where the second equality follows from $\hat{\mathbb{E}}\hat{u} = 0$, which in turn follows by construction of the sample linear regression, and where k is the number of regressors.
    - Note that the sample variance of $\hat{u}$ is given by applying an adjustment analogous to Bessel's correction to the sample analogue of variance. This adjustment is necessary, informally, because each of the sample linear regression parameters is constructed to minimise the residual, hence the variance of the residual in the sample linear regression is systematically less than the variance of the residual in the population linear regression, so some correction is necessary to eliminate this bias.
  - The standard error of regression is $s_{\hat{u}} := \sqrt{s_{\hat{u}}^2}$.

R^2

- The standard error of regression gives an absolute measure of the variability of Y_i around the fitted (predicted) value $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots$, the R^2 of a regression gives a relative measure of this variability.
- The total sum of squares of some sample linear regression is $TSS := \sum_{i=1}^{n}(Y_i - \hat{\mathbb{E}}Y)^2 = n\hat{\mathbb{E}}(Y - \hat{\mathbb{E}}Y)^2$. The explained sum of squares is $ESS := \sum_{i=1}^{n}(\hat{Y}_i - \hat{\mathbb{E}}Y)^2 = n\hat{\mathbb{E}}(\hat{Y} - \hat{\mathbb{E}}Y)^2$. The sum of squared residuals is $SSR := \sum_{i=1}^{n}\hat{u}_i^2 = n\hat{\mathbb{E}}\hat{u}$.

$$
\begin{aligned}
TSS := \sum_{i=1}^{n}(Y_i - \hat{\mathbb{E}}Y)^2 &= \sum_{i=1}^{n}[\hat{Y}_i + \hat{u}_i - \hat{\mathbb{E}}Y]^2 \\
&= \sum_{i=1}^{n}[(\hat{Y}_i - \hat{\mathbb{E}}Y) - \hat{u}_i]^2 \\
&= \sum_{i=1}^{n}[(\hat{Y}_i - \hat{\mathbb{E}}Y)^2 + \hat{u}_i^2 - 2(\hat{Y}_i - \hat{\mathbb{E}}Y)\hat{u}_i] \\
&= \sum_{i=1}^{n}(\hat{Y}_i - \hat{\mathbb{E}}Y)^2 + \sum_{i=1}^{n}\hat{u}_i^2 - 2\sum_{i=1}^{n}(\hat{Y}_i - \hat{\mathbb{E}}Y)\hat{u}_i \\
&= \sum_{i=1}^{n}(\hat{Y}_i - \hat{\mathbb{E}}Y)^2 + \sum_{i=1}^{n}\hat{u}_i^2 - 2n\hat{\mathbb{E}}(\hat{Y}_i - \hat{\mathbb{E}}Y)\hat{u} \\
&= \sum_{i=1}^{n}(\hat{Y}_i - \hat{\mathbb{E}}Y)^2 + \sum_{i=1}^{n}\hat{u}_i^2 - 2n[\hat{\mathbb{E}}\hat{Y}\hat{u} - \hat{\mathbb{E}}\hat{Y}\hat{\mathbb{E}}\hat{u}] \\
&= \sum_{i=1}^{n}(\hat{Y}_i - \hat{\mathbb{E}}Y)^2 + \sum_{i=1}^{n}\hat{u}_i^2 \\
&= ESS + SSR
\end{aligned}
$$

  - where the seventh equality follows from the fact that $\hat{Y}$ is a linear combination of $Y, X_1, X_2, \ldots$, each of which is uncorrelated with $\hat{u}$, by construction of the sample linear regression parameters.
- $R^2 := \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$.
  - Intuitively, the R^2 of a regression is the proportion of the total variability in the outcome attributable to variation in the regressors.
- Note that the standard error $s_{\hat{u}} = (\frac{SSR}{n-k-1})^{\frac{1}{2}}$.
- The sample linear regression parameters minimise SSR, then inclusion of an additional regressor results in a weak decrease in SSR, because the earlier SSR can be obtained by assigning 0 to the parameter of the additional regressor. Including irrelevant regressors improves fit, as measured by SSR, but undermines fit, as measured by $s_{\hat{u}}$ because of the degrees of freedom correction.
- The adjusted R^2 is $\overline{R}^2 := 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$.
  - This adjustment can be understood as applying two simultaneous adjustments, one to TSS and the other to SSR. TSS is adjusted by some analogue of Bessel's correction, and SSR is adjusted by some analogue of the degrees of freedom adjustment. Then, the adjusted TSS is an unbiased estimator of $var(Y)$ (in the population) and the adjusted SSR is an unbiased estimator of $var(u)$ (in the population).

Inference on Regression Parameters

- In what follows, it is supposed that:
  - (1) the causal model satisfies orthogonality,
  - (2) the sample is an independently and identically distributed random sample,
  - (3) large outliers are unlikely, and
    - This is measured by the kurtosis, which is the expectation of the fourth power, standard normal has kurtosis of 3, so excess kurtosis is measured by subtracting 3. For the central limit theorem to apply, the kurtosis needs to be finite.
  - (4) there is no perfect multicollinearity among any subset of the regressors.
- The law of large numbers implies that the sample linear regression parameters are consistent estimators of the parameters of the causal model.
- The central limit theorem implies that $\frac{\hat{\beta}_I - \beta_I}{se(\hat{\beta}_I)} \approx \frac{\hat{\beta}_I - \beta_I}{sd(\hat{\beta}_I)} \to^d N[0,1]$, for arbitrary regressor with index I.
- Let ω_{β_I} and $\hat{\omega}_{\beta_I}$ be such that $sd(\hat{\beta}_I) = n^{-\frac{1}{2}}\omega_{\beta_I}$ and $se(\hat{\beta}_I) = n^{-\frac{1}{2}}\hat{\omega}_{\beta_I}$.
  - It can be proven that $\omega_{\beta_I}^2 = \frac{\mathbb{E}(\tilde{X}_I - \mathbb{E}\tilde{X}_I)^2u^2}{[\mathbb{E}(\tilde{X}_I - \mathbb{E}\tilde{X}_I)^2]^2} = \frac{\mathbb{E}\tilde{X}_I^2u^2}{[\mathbb{E}\tilde{X}_I^2]^2}$.
    - The second equality follows from $\mathbb{E}\tilde{X}_I = 0$ and $\mathbb{E}\tilde{X}_Iu = 0$. The former holds by construction of the (population) linear regression of X_I on the remaining regressors. The latter holds because $\tilde{X}_I$ is a linear function of the remaining regressors, each of which is uncorrelated with u, by construction of the (population) linear regression of Y on $X_1, X_2, \ldots$.
  - The residual u in the population regression model is homoskedastic iff $\mathbb{E}[u|X_1, X_2, \ldots] = 0$ and $\mathbb{E}[u^2|X_1, X_2, \ldots] = \mathbb{E}u^2$, where $u, X_1, X_2, \ldots$ are drawn from some arbitrary observation in the population.
  - Supposing that u is homoskedastic,

$$\begin{aligned}\mathbb{E}\tilde{X}_1^2u_1^2 &= \mathbb{E}(\mathbb{E}[\tilde{X}_I^2u^2|X_1,X_2,\ldots]) \\ &= \mathbb{E}(\tilde{X}_I^2\mathbb{E}[u^2|X_1,X_2,\ldots]) \\ &= \mathbb{E}u^2\mathbb{E}\tilde{X}_1^2 \\ &= \sigma_u^2 var(\tilde{X}_I)\end{aligned},$$

    - where the first equality holds by the law of iterated expectations, the second equality holds by conditioning (because within the conditional, $\tilde{X}_I$ is known), the third equality holds by homoskedasticity, and the fourth equality holds by definition, and the fact that $\mathbb{E}u = \mathbb{E}\tilde{X}_I = 0$.
  - Then, $\omega_{\beta_I} = \frac{\sigma_u^2}{var(\tilde{X}_I)}$.
  - The standard deviation of β_I decreases with decreasing variability of the residual in the population linear regression model, and with increasing variability of the part of X_I not correlated with the other regressors.
- Then, supposing that u is homoskedastic, the standard error for $\hat{\beta}_I$ is computed as $se(\hat{\beta}_I) = n^{-\frac{1}{2}}\sqrt{\frac{s_u^2}{\hat{var}(\hat{X}_1)}} = n^{-\frac{1}{2}}\frac{s_{\hat{u}}}{\hat{sd}(\hat{X}_1)}$. If homoskedasticity is not a plausible assumption, the standard error for $\hat{\beta}_I$ is computed as $se(\hat{\beta}_I) = n^{-\frac{1}{2}}\sqrt{\frac{\hat{var}(\hat{\tilde{X}}_I\hat{u})}{[\hat{var}(\hat{X}_I]^2}} = n^{-\frac{1}{2}}\frac{\hat{sd}(\hat{\tilde{X}}_1\hat{u})}{\hat{var}(\hat{X}_1)}$.
- The t-statistic is $t(b) := \frac{\hat{\beta}_I - b}{se(\hat{\beta}_I)}$, where b is the hypothesised value of the parameter of interest β_I.
- The distribution of the test statistic under the null, and the decision rule, are exactly analogous to ordinary hypothesis tests.
- The p-value is exactly analogous to ordinary hypothesis tests.
- The confidence interval is exactly analogous to ordinary hypothesis tests.
- In general, removing a regressor has an ambiguous effect on the precision (measured by the standard error $se(\hat{\beta}_I)$) of the remaining regressors because the variation of the residual decreases but the variation of the residual of the Frisch-Waugh-Lovell auxiliary regression also decreases (because a regressor is removed from here too).

F Tests

- The basic idea behind the F test (supposing homoskedasticity) is to infer from the deterioration of the fit of a linear regression model, measured in terms of SSR or R^2, as a consequence of imposing the null hypothesis, the plausibility of the null hypothesis.
- Let $\hat{u}_{rs,i}$ denote the residual of the i^{th} observation in the restricted sample linear regression, and $\hat{u}_{un,i}$ denote the residual of the i^{th} observation in the unrestricted sample linear regression. Then the respective sum of squared residuals are $SSR_{rs} = \sum_{i=1}^n \hat{u}_{rs,i}^2$ and $SSR_{un} = \sum_{i=1}^n \hat{u}_{un,i}^2$, and the difference is denoted by $\Delta = SSR_{rs} - SSR_{un}$.
  - Δ is necessarily non-negative because the regressions are constructed to minimise the respective SSRs, and the parameters of the restricted regression are "feasible" in the unrestricted regression, so SSR_{un} must be weakly less than SSR_{rs}.
  - Intuitively, the larger Δ, the greater the deterioration in the fit of the model upon imposing the restrictions entailed by the null.
  - The magnitude of Δ is dependent on the scale of the data, so it is appropriate to adjust for the scale of the data by dividing Δ by SSR_{un}. SSR_{un} increases directly proportionately to the number of degrees of freedom $n-k-1$, so a further adjustment to eliminate this effect is appropriate. The resulting statistic is the likelihood ratio test statistic $LR := \frac{SSR_{rs}-SSR_{un}}{\frac{SSR_{un}}{n-k-1}}$.
  - Under the null, $LR \to^d \chi_q^2$, where χ_q^2 is the chi-squared distribution with q degrees of freedom, where q is the number of restrictions imposed by the null. This implies $F := \frac{LR}{q} \to^d \frac{\chi_q^2}{q} = F_{q,\infty}$, where $F_{q,\infty}$ is an F distribution with q numerator and ∞ denominator.
    - $F_{q,\infty} = \frac{1}{q}\chi_q^2 = \frac{1}{q}\sum_{i=1}^q Z^2$.
  - Noting that $1 - R_m^2 = \frac{SSR_m}{TSS}$ for $m \in \{un, rs\}$, and that TSS is model-independent (it is entirely a function of the data), the F statistic can be equivalently stated as $F = \frac{R_{un}^2 - R_{rs}^2}{1-R_{un}^2} / \frac{q}{n-k-1}$.
- A special case is the test of the hypothesis that none of the regressors is relevant. In this case, the number of restrictions q is equal to the number of regressors k, $SSR_{rs} = TSS$, and $R_{rs}^2 = 0$.
- (See Duffy, 2022 - Notes on Linear Regression 2, pp. 43-44 for an application of F tests to more general linear restrictions).

Proxies

- An appropriate proxy is one that is (1) correlated with the unobserved determinant that it serves as a proxy for, (2) uncorrelated with the other unobserved determinants, such that orthogonality holds, and (3) such that the prediction of the proxied determinant is not improved by knowledge of the other observed determinants.
  - Consider the causal model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$. Suppose that X_2 is unobserved. Consider the population regression model $X_2 = \pi_0 + \pi_1 X_1 + \pi_2 W_2 + v$, where W_2 is a proxy for X_2 and $\mathbb{E}v = \mathbb{E}X_1 v = \mathbb{E}W_2 v = 0$. Then, by substitution $Y = (\beta_0 + \beta_2 \pi_0) + (\beta_1 + \beta_2 \pi_1) X_1 + \beta_2 \pi_2 W_2 + (u + \beta_2 v)$. Supposing that (3) holds, $\pi_1 = 0$, the above reduces to $Y = (\beta_0 + \beta_2 \pi_0) + \beta_1 X_1 + \beta_2 \pi_2 W_2 + (u + \beta_2 v)$. Supposing that (2) holds, orthogonality holds in the causal model, then its coefficients can be consistently estimated by OLS regression. In particular, β_1 can be consistently estimated by OLS regression.

Perfect Multicollinearity

- Parameters in a causal model that includes perfectly multicollinear determinants cannot be interpreted as the causal effect of each determinant because a change in one of the perfectly multicollinear determinants is necessarily accompanied by a change in at least one of the other perfectly multicollinear determinants. There is no (in the reality that is being modelled) variation in (only) one of the perfectly multicollinear determinants. It does not make sense to speak of a variation in one of the perfectly multicollinear determinants while holding the others constant.
- Parameters in a causal model that includes perfectly multicollinear determinants cannot be estimated (at all, let alone consistently) by OLS regression because the residual in each Frisch-Waugh-Lovell auxiliary regression is zero and has zero variance, hence the OLS coefficient (which has this variance in its denominator) is undefined. There is no unique solution to the OLS regression problem.
- Perfect multicollinearity could remain in the data even if one of the perfectly multicollinear (in theory) regressors is omitted. This is the case if one of the remaining regressors is a constant in the sample.