Interpreting regression output

Contents

Disclaimer: Sections and lines in brown correspond to content which is very much 'under construction'.

1 Elements of the typical linear regression output summary

```
> mod1 = lm(dist ~ speed, data = cars)
> summary(mod1)

Residuals:
     Min       1Q   Median       3Q      Max
-29.069   -9.525   -2.272    9.215   43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.9800     2.1750  19.761  < 2e-16 ***
speed.c       3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12
```

Figure 1: Summary of the results of the OLS estimation of a univariate linear regression model, in `R`

> Regression is a tool for estimating average differences across groups. We estimate the difference in average observed outcomes.

Residuals $\{r_i\}_i$ Difference between the observed response values y_i and those predicted $\hat{y}_i$.
Errors $\varepsilon_i = y_i - X_i\beta$, residuals are estimates of the errors: $r_i = y_i - X_i\hat{\beta} = y_i - \hat{y}_i$
Plot them to look at their distribution: is it centered around 0, is it normal...?

Coefficients For each coefficient, an estimate and the level of uncertainty for that estimate.

- **Slope estimate** $\hat{\beta}_j^{\text{OLS}} \equiv \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} r_i^2 = (X'X)^{-1}X'y$

  Interpretation as comparison:

  - Slope coefficients should be interpreted as *comparisons between units* that differ in one predictor: *"$\hat{\beta}_j$ = how y differs, on average, when comparing units that differ by 1 in the predictor x_j."*[1]
  - With multiple predictors, interpretations are contingent on the other variables in the model. → Interpret each coefficient "with all the other predictors held constant". Note also that if predictors are correlated, $\hat{\beta}_j$ measures not the *total* change in y associated with a difference in x_j, but the *additional* change associated with the change in x_j, when the effects of all other variables are already accounted for.
  - By nature of x_j:

[1] Interpretation as changes *within* units, i.e., a "counterfactual" interpretation as *the expected change in y caused by adding 1 to x_y* requires a causal identification. I.e., from the data alone, a regression only tells us about comparisons between units, not about changes within units.

* continuous: $\hat{\beta}_j$ = average difference in y for a 1-unit difference in x_j, holding other x's constant;
* categorical, e.g., binary: $\hat{\beta}_j$ = average difference in y between the category for which $x_j = 0$ and the category for which = 1.

If there is a single regressor, the estimand is $\beta_{\text{OLS}} = \frac{\text{cov}[x,y]}{\mathbb{V}[x]}$ and the estimator (its sample analog) $\hat{\beta}_{\text{OLS}} = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_i (x_i - \bar{x})^2}$. In the multivariate regression model, each estimand is $\beta^k_{\text{OLS}} = \frac{\text{cov}[\tilde{x}_k, y]}{\mathbb{V}[\tilde{x}_k]}$ where $\tilde{x}_k$ is the residual from the regression of x_k on all the other covariates.

Centering or standardizing X:

- Centering = subtracting the mean from each column of X: it only affects the intercept, and allows us to interpret it as the expectation of y in a linear model.
- Standardizing = subtracting the mean of each predictor and dividing by the standard deviation: puts all predictors on a common scale. $\hat{\beta}_j$ is the expected difference in y, comparing units that differ by one standard deviation in x_j, with all other predictors fixed at their average values.

- **(estimated) Standard Error $\text{SE}[\hat{\beta}_j] = \dfrac{\hat{\sigma}_{\hat{\beta}_j}}{\sqrt{n}}$**

  = an estimate of the standard deviation of $\hat{\beta}_j$'s sampling distribution.
  The SE gives us a sense of our uncertainty about $\hat{\beta}_j$: the expected difference in $\hat{\beta}_j$ if we were to run the model again and again. A lower SE *relative to the coefficient* means more certainty.
  SEs are used in computing confidence intervals and in the t-statistic for hypothesis testing.

- **t-statistic $t_{\hat{\beta}} = \dfrac{\hat{\beta} - \beta_0}{\text{SE}[\hat{\beta}]}$**

  The realization of the t-statistic for the null hypothesis H_0: $\beta_j = 0$.
  t_β can be used in a two-sided[2] t-test of H_0, as, *if the error term is normally distributed*, it follows a Student's t-distribution under H_0: $t_\beta \underset{H_0}{\sim} \mathcal{T}_{n-2}$. If its realization $t_{\hat{\beta}}$ falls in the tails of that distribution, that would mean it is very unlikely given H_0, therefore we can reject H_0. We will examine that with the two-sided p-value.

- **p-value = $\Pr\left(\text{observing a } T > |t_{\hat{\beta}}|\right)$ under H_0**

  I.e., the probability of observing data as extreme as that actually observed, assuming H_0.[3]

  p-value small (< 0.05) $\iff$ $t_{\hat{\beta}}$ falls in the tail of the Student's $\mathcal{T}$-distribution
  $\implies$ observing our $t_{\hat{\beta}}$ is highly unlikely under H_0
  $\implies$ reject H_0
  $\implies$ there is a relationship between y and x, $\hat{\beta}$ is "significant".

Residual Standard Error or **Standard Error of the Regression (SER)**
Summary of the scale of the residuals. It is the average distance by which an observed value falls from the

[2] By default, statistical packages carry out a two-sided test and therefore report the two-sided p-value; however we could also use the t-statistic to carry out a one-sided test.

[3] ⚠ The p-value is often misinterpreted to be the probability that H_0 is true, when it is the probability of observing data as extreme or more extreme than that actually observed, assuming H_0. *p-value* = P(obs | hyp) ≠ P(hyp | obs)

regression line, i.e., the accuracy to which the model can predict y. Interpretations:

- *y can deviate from the true regression line by 15.38 ft, on average.*
- *The model can predict y to an accuracy of about 15.38 points.*
- *About 68% of y will be within 15.38 of the predicted value.*
- *SER^2 = the variance "unexplained" by the model: the amount of variation remaining in y after we remove the variation due to x.*

R^2 or **coefficient of determination**
How much better the sample data is fit by the sample regression line $y = \alpha + \beta X$ than by the sample mean line, $y = \bar{Y}$. It is one way of measuring the goodness-of-fit. See Figure 2.

$$\mathrm{R}^2 = 1 - \frac{\text{sum of squared residuals (SSR)}}{\text{total sum of squares (TSS)}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad \in [0, 1]$$

– In most linear models, $TSS = ESS + RSS$, where ESS is the explained sum of squares $\sum_i (\hat{y}_i - \bar{y})^2$, therefore $\mathrm{R}^2 = \frac{ESS}{TSS}$ = the proportion of the sample variance in y that is explained by X.

– ⚠ R^2 mechanically increases as more predictors are included in the regression. → Use the **adjusted R^2** which adjusts for the number of predictors.

F statistic for an F-test[4] of overall significance.
H_0: all coefficients equal 0 (no relationship between y and X). H_a: $\beta_j \neq 0$, for at least one j. We test the full model against a model with no regressors.
F's p-value is small $\iff$ at least some of the parameters are nonzero and the regression equation does have some validity in fitting the data (the X's are not purely random w.r.t. y)

$$F = \frac{\text{mean regression sum of squares (MSR)}}{\text{mean error sum of squares (MSE)}} = \frac{\frac{ESS}{k}}{\frac{SSR}{n-k-1}} \quad \in [0, +\infty[$$

[4] In general, an F-statistic is a ratio of two quantities, F-test tests the H_0 that the quantities are roughly equal, i.e. $F \simeq 1$. Reject H_0 if F high ($>> 1$). How large F really needs to be depends on the number of data points and predictors: if large sample, F-stat slightly above 1 is sufficient to reject H_0; if small sample, need a large F-stat.

2 Common test statistics for regression models

We consider tests used to assess the statistical significance of explanatory variables, in a regression of y on p covariates X. The general approach to conducting a statistical test consists of the following steps:

1. Write the null hypothesis H_0 — the hypothesis to nullify.
2. Design a test statistic T that summarizes the data's deviation from what would be expected under H_0, and that has a specific distribution under H_0. Ex:
   - an F-test is any test in which the test statistic has an $\mathcal{F}$ distribution under H_0.
   - a t-test is any test in which the test statistic has a Student's $\mathcal{T}$ distribution under H_0;
   - a Wald test is a test in which the test statistic has an *asymptotic* χ^2 distribution under H_0;
   - a z-test is any test in which the test statistic has an *approximately* normal distribution under H_0.
3. Compute the realized value of T for our data: T_{obs}.
4. If it falls in the tails of the distribution, i.e., it is very unlikely given H_0, we can reasonably reject H_0.

H_0	Test statistic T	Inference
z-test and t-test *Is the coefficient β_j significant?* H_0: $\beta_j = 0$	$z \equiv \dfrac{\hat{\beta} - \beta_0}{\text{SD}[\hat{\beta}]}$ $t \equiv \dfrac{\hat{\beta} - \beta_0}{\widehat{\text{SD}}[\hat{\beta}]} = \dfrac{\hat{\beta} - \beta_0}{\text{SE}[\hat{\beta}]}$	IF { assumptions on means, covariances... } $\implies t \overset{a}{\underset{H_0}{\sim}} \mathcal{N}(0,1)$ IF { iid normal data } $\implies t \underset{H_0}{\sim} t_{n-p}$ $z \underset{H_0}{\sim} \mathcal{N}(0,1)$
Wald Chi-squared test *Are the k coefficients β jointly significant?* H_0 : $\beta = 0_{k\times 1}$, for k parameters among the $p+1$[5]	$W \equiv (\hat{\beta} - \beta_0)' \, \hat{\mathbb{V}}_{\hat{\beta}}^{-1} (\hat{\beta} - \beta_0)$	with MLE estimates $\implies W \overset{a}{\underset{H_0}{\sim}} \chi^2_k$ IF { iid normal data } $\implies \frac{W}{k} \underset{H_0}{\sim} \mathcal{F}_{k,n-k}$ ($W \underset{H_0}{\sim} \chi^2_k$ if σ^2 known)
F-test of overall significance *Is our model a significantly better fit than one with no predictors?* H_0 : $\beta = 0_{p\times 1}$, for all p slope parameters	$F \equiv$ [6] $\dfrac{\text{ESS}/(p-1)}{\text{RSS}/(n-p)}$ $= \dfrac{\sum_i (\hat{y}_i - \bar{y})^2/(p-1)}{\sum_i (y_i - \hat{y}_i)^2/(n-p)}$	IF – the scaled SSs are $\perp\!\!\!\perp$ – each SS $\sim \chi^2$ (guaranteed if iid normal data) $\implies F \underset{H_0}{\sim} \mathcal{F}_{p-1,n-p}$

The asymptotic standard normal distributions stem from the slope parameters having asymptotic normal distributions themselves (from the CLT). So *for large n only*, we can say that the test statistics are approximately normally distributed. However, for small n, they are close to normal only if the data themselves are normal. E.g., if the data are substantially non-normal, or in the presence of outliers, the t-test can give misleading results.

[5] If $p = 1$, the Wald statistic is equivalent to the squared t statistic: $W \equiv \frac{(\hat{\beta}-\beta_0)^2}{\hat{\mathbb{V}}_{\hat{\beta}}}$, $\chi^2_1 = \mathcal{N}(0,1)^2$, $\mathcal{F}_{1,n-1} = t^2_{n-1}$.

[6] The statistic is the ratio of the explained variance to the unexplained variance. ESS is the explained sum of squares, RSS the residual sum of squares.

3 Elements of the typical logistic regression output summary

- Log likelihood of the model. This value has no meaning in and of itself; rather, this number can be used to help compare nested models.
- Pseudo R2: Logistic regression does not have an equivalent to the R-squared that is found in OLS regression; however, many people have tried to come up with one. There are a wide variety of pseudo-R-square statistics. Because this statistic does not mean what R-square means in OLS regression (the proportion of variance explained by the predictors), we suggest interpreting this statistic with great caution.
- Coef: They are usually by default in log-odds units, which are often difficult to interpret, so they may be converted into odds ratios. You can do this by hand by exponentiating the coefficient, or by using the or option with logit command, or by using the logistic command.

a. Test each coefficient

The test statistics to assess the significance of the regression parameters in logistic regression analysis are based on chi-square statistics, as opposed to t statistics as was the case with linear regression analysis. This is because we use a different estimation technique: MLE.

For each regression coefficient of the predictors, we can use a z-test (note not the t-test). In the output, we have z-values and corresponding p-values.

b. Test the overall model

For the linear regression, we evaluate the overall model fit by looking at the variance explained by all the predictors. For the logistic regression, we cannot calculate a variance. However, we can evaluate the deviance. For a model without any predictor, we can calculate a null deviance, which is similar to variance for the normal outcome variable. After including the predictors, we have the residual deviance. The difference between the null deviance and the residual deviance tells how much the predictors help predict the outcome. If the difference is significant, then overall, the predictors are significant statistically.

The difference or the decease in deviance after including the predictors follows a chi-square distribution. (Rmk: It has a close relationship to F distribution: it is the limiting distribution of an F distribution as the denominator degrees of freedom goes to infinity.)

There are two ways to conduct the test. From the output, we can find the Null and Residual deviances and the corresponding degrees of freedom. Then we calculate the difference. For the mammography example, we first get the difference between the Null deviance and the Residual deviance, 203.32-155.48= 47.84. Then, we find the difference in the degrees of freedom 163-159=4. Then, the p-value can be calculated based on a chi-square distribution with the degree of freedom 4. Because the p-value is smaller than 0.05, the overall model is significant.

4 Transformations

Inverse hyperbolic sine (IHS) transformation For outcomes that have a thick right tail, the standard solution is to take a log transformation;[7] it brings extreme values closer to the middle, so they don't have such a large effect on the results. However, when the outcome also has many zero-valued observations (e.g., wealth), natural log transformations don't work well as $\ln(0)$ is undefined.

Instead one can use the inverse hyperbolic sine (IHS or arcsinh) transformation:

$$\log\left(y_i + (y_i^2 + 1)^{\frac{1}{2}}\right)$$

It approximates the natural logarithm (except for very small values of y, it is $\approx \log(2y_i) = \log(2) + \log(y_i)$), and so it can be interpreted in exactly the same way as a standard log-transformed dependent variable, but is defined at zero, thus allows retaining zero-valued observations.

[7] Another solution is to run quantile regressions and analyze each part of the distribution separately.

5 Interpreting coefficients of a regression with...

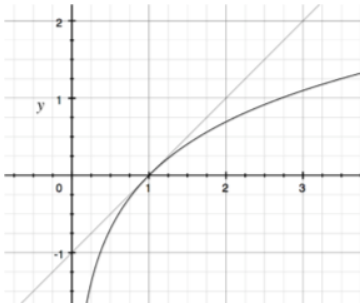
5.1 ... Log transformations

Regression model		*Given a change in x, what change do we expect in y?*
level-level (linear)	$y = \beta_0 + \beta_1 x + e$	If x increases by 1 unit, y increases by β_1 units.
log-level (log-linear)	$\ln(y) = \beta_0 + \beta_1 x + e$	If x increases by 1 unit, $\ln(y)$ increases by β_1 units, i.e. y increases by a factor e^{β_1}. • if small $\hat{\beta}$: can approximate: y increases by $(100 \times \beta_1)\%$ • if large $\hat{\beta}$: approximation invalid. y increases by $\times e^{\beta_1}$
level-log	$y = \beta_0 + \beta_1 \ln(x) + e$	If $\ln(x)$ increases by 1 unit, y increases by β_1 units, i.e. if x increases by 1%, y increases by $\frac{\beta_1}{100}$ units.
log-log	$\ln(y) = \beta_0 + \beta_1 \ln(x) + e$	If x increases by 1%, y increases by β_1%. (β_1 is an *elasticity*.)

Why can we interpret natural log changes as percentage changes?

The log function is approximately linear around 1, i.e., it is reasonable to do a first order Taylor approximation of $\ln(x)$ around $x = 1$:

$$f(x) \simeq f'(1)(x-1) + f(1)$$
$$\ln(x) \simeq \tfrac{1}{1}(x-1) + 0$$
$$\ln(x) \simeq x - 1$$



Then a *small* difference in logs of y can be approximately interpreted as a percentage change in y:

$$\ln(y_2) - \ln(y_1) = \ln\left(\frac{y_2}{y_1}\right) \simeq \frac{y_2}{y_1} - 1 = \frac{y_2 - y_1}{y_1}$$

Therefore, in a log-linear regression $\ln(y) = \beta_0 + \beta_1 x + e$:

– if $\hat{\beta}_1$ small, one can say *"A 1-unit increase in x corresponds to a $(100 \times \hat{\beta}_1)\%$ increase in y"*;
– if $\hat{\beta}_1$ large, one should stick to saying *"A 1-unit increase in x corresponds to a $e^{\hat{\beta}_1}$ factor increase in y"*.

Note: A classical approach for regression modeling of a right-skewed outcome on the positive real line is to log-transform the outcome, $\log(Y)$. Indeed, if the outcome is highly skewed, the log transformation effectively "pulls in" high Y values that appear in the upper tail of a right-skewed distribution, narrowing its range. It will make the outcome normal enough that linear regression is valid. When we apply standard linear regression, we are then making the implicit assumption that $\log(Y)$ is normally distributed: $\log(Y) \sim \mathcal{N}(\mu, \sigma^2)$. With covariates, the model is: $\log(Y|X_1, ..., X_k) \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k, \sigma^2)$.

5.2 ... Interacted predictors

Adding an interaction term allows the slope to vary across subgroups, and changes the interpretation of all coefficients along the way. Examples:

- kid_score = $\beta_0 + \beta_1$ mom_hs + β_2 mom_iq + β_3 mom_hs : mom_iq
  β_3 represents the difference in the slope for *mom_iq*, comparing children with mothers who did and did not complete high school.
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ The effect of x_1 is $\beta_1 + \beta_3 x_2$: it is different for each value of x_2.

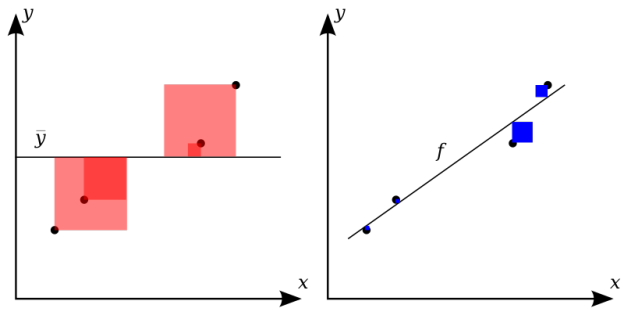


Figure 2: Representation of the terms of the coefficient of determination $R^2 = 1 - \frac{RSS}{TSS}$. The red areas represent the squared residuals w.r.t. to the average value $\bar{y}$, the blue areas represent the squared residuals w.r.t. the linear regression. The better the linear regression fits the data in comparison to the simple average, the higher the R^2.

Source: Orzetto - https://commons.wikimedia.org/w/index.php?curid=11398293