# Interpreting regression output

## Contents



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

# 1 Elements of the typical 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 $\hat{\beta}_j$**
  - With multiple predictors, the interpretation of any given coefficient is, in part, contingent on the other variables in the model. $\rightarrow$ Interpret each coefficient "with all the other predictors held constant".
    Furthermore, if predictors are correlated, it is important to note the following distinction: $\hat{\beta}_j$ measures not the *total* change in $y$ expected from increasing $x_j$, but the *additional* change from increasing $x_j$, when the effects of all other variables are already accounted for.
  - **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 two groups of units that differ by 1 in the predictor $x_j$."* Interpretation as changes *within* units, i.e. a "counterfactual" interpretation *"$\hat{\beta}_j$ = the expected change in $y$ caused by adding 1 to $x_y$"* requires justification other than the data, e.g., in causal inference studies.
  - $x_j$ continuous: $\hat{\beta}_j$ = average change in $y$ for a 1-unit change in $x_j$, holding other $x$'s constant

– $x_j$ categorical: e.g. binary: $\hat{\beta}_j$ = average difference in $y$ between the category for which $x_j = 0$ and the category for which $= 1$

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

If there is a single regressor, $\hat{\beta} = \dfrac{\text{cov}[x, y]}{\mathbb{V}[x]}$

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

  = 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** $t_{\hat{\boldsymbol{\beta}}} = \dfrac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0}{\widehat{\mathbf{SE}}(\hat{\boldsymbol{\beta}})}$
  The realization of the t-statistic for the null hypothesis $H_0$: $\beta_j = 0$.
  $t_\beta$ can be used in a two-sided[1] 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** $= \mathbf{Pr}\Big($**observing a** $\boldsymbol{T} > |\boldsymbol{t}_{\hat{\boldsymbol{\beta}}}|\Big)$ **under** $\boldsymbol{H_0}$
  I.e., the probability of observing data as extreme as that actually observed, assuming $H_0$.[2]

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

### 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 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

[1] 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.

[2] ⚠ 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)

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 $\boldsymbol{R^2}$** which adjusts for the number of predictors.

**F statistic** for an F-test[3] 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 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. $y$)

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

[3] 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 Transformations

**Inverse hyperbolic sine (IHS) transformation** For outcomes that have a thick right tail, the standard solution is to take a log transformation[4]; 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.

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

# 3 Interpreting coefficients of a regression with...

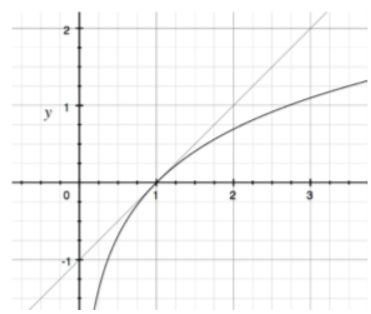
## 3.1 ... Log transformations

| 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 $\beta_1$ units. |
| log-level (log-linear) | $\ln(y) = \beta_0 + \beta_1 x + e$ | If $x$ increases by 1 unit, $\ln(y)$ increases by $\beta_1$ units, i.e. $y$ increases by a factor $e^{\beta_1}$. <br>• if small $\hat{\beta}$: can approximate: $y$ increases by $(100 \times \beta_1)\%$ <br>• 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 $\beta_1$ 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 $\beta_1$%. ($\beta_1$ 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 $x = 1$:

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




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"*.

## 3.2 ... 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 $+ \beta_2$ mom_iq $+ \beta_3$ mom_hs : mom_iq
  $\beta_3$ 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 $x_2$.

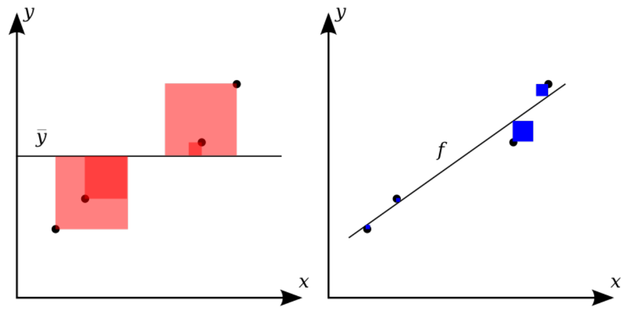



Figure 2: Representation of the terms of the coefficient of determination $R^2 = 1 - \frac{SSR}{TSS}$. The red areas represent the squared residuals w.r.t. to the average value $\bar{y}$, the blue areas represent the squared residuals w.r.t. the linear regression. The better the linear regression fits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

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