Exercise Set 11 - Solution

1 Confidence Intervals [basic]

We assume that the polymers' strength is a random variable following the normal distribution. Since we don't know its variance in the population, we have to estimate it using the sample variance.

$$\bar{y} = 10.665 \qquad s_y^2 = 12.077$$

We have $n = 10$ data points, thus we know that the T-statistic will follow a $t_{n-1} = t_9$ distribution. The value of t is given by:

$$t = \frac{\bar{y} - \mu_0}{s_y/\sqrt{n}}$$

(Remember that the population variance is s_y^2 but the variance of the mean of n data points is s_y^2/n.) A level of α means that the fraction $1 - \alpha$ of that distribution should be within the α confidence interval. Here we are doing a two-sided test, so to get a confidence level of $1 - \alpha$ we need to compute $2 \times \mathrm{CDF}(t) - 1 = 1 - \alpha$, so for a significance level of $\alpha = 0.1 = 10\%$ we need to use the value of t where $\mathrm{CDF}(t) = 0.95$ i.e. the $qt_9(95\%)$ quantile, which is 1.833. Using the quantile form the student-test table (e.g. on www.tdistributiontable.com), we obtain:

$$y \in \left[\bar{y} \pm qt_9(95\%)\frac{s_y}{\sqrt{N_s}}\right] = [8.65; 12.68]$$

I.e. there is a 90% probability that the true mean is within this range, and only a 10% probability that it is outside this range.

For other α, the intervals are:

$$\begin{aligned}
\alpha = 0.05 \qquad [\bar{y} \pm qt_9(97.5\%)\frac{s_y}{\sqrt{n}}] &= [8.18; 13.15] \\
\alpha = 0.01 \qquad [\bar{y} \pm qt_9(99.5\%)\frac{s_y}{\sqrt{n}}] &= [7.09; 14.24] \\
\alpha = 0.001 \quad [\bar{y} \pm qt_9(99.95\%)\frac{s_y}{\sqrt{n}}] &= [5.41; 15.92]
\end{aligned}$$

We observe that the size of the interval increases as the significance level decreases. This makes sense: We use the significance to mean "only with a probability of α could random fluctuations cause the mean be outside the confidence interval" (where the latter part of the sentence is the null hypothesis we test). Inversely speaking, if the null hypothesis is true, this means that we expect the mean to be within the confidence interval with probability $1-\alpha$. The bigger the interval, the bigger our confidence to find the mean in that range.

In the extreme, take $\alpha = 0$. Then the confidence interval would be $[-\infty; \infty]$ - as there is always some (very small) probability for *any* possible value of $\bar{y}$ to occur randomly. Of course as long as the approximation of $\bar{y}$ coming from a normal distribution is still correct - often so far away from the mean this may fail. For example the mean height of a group of people may be well described by a normal distribution for values reasonably close to the mean. But this can obviously not be true for any distance from the mean, as then we would have to give a non-zero probability to negative heights!

The sample dataset was generated using a normal distribution $\mathcal{N}(11, 9)$, i.e. with a true population mean of 11, which indeed falls into every confidence interval computed before.

2 Variance of intercept estimator [normal]

From expression of the regression, the intercept can be rewritten as $\hat{a} = \bar{Y} - \hat{b}\bar{X}$. So:

$$\mathbb{V}(\hat{a}) = \mathbb{V}(\bar{Y}) + \mathbb{V}(\hat{b}\bar{X}^2) = \mathbb{V}(\bar{Y}) + \bar{X}^2 \cdot \mathbb{V}(\hat{b}) = \frac{\sigma^2}{n} + \bar{X}^2 \cdot \frac{\sigma^2}{\sum_i (x_i - \bar{X})^2}$$

To get this, note that for two independent variables A and B, we have $\mathbb{V}(A - B) = \mathbb{V}(A) + \mathbb{V}(B)$. Although X and Y may be correlated, the mean of Y is not correlated to the mean of X, as the latter is not varying. Simple linear regression is based on the assumption that X does not have any error (and hence the mean of X has no variance). Also the mean of Y does not correlate with the slope estimator $\hat{b}$ (as can be seen from the definition of $\hat{b}$). Finally, the expression for $\mathbb{V}(\bar{Y})$ more generally follows from the Central Limit Theorem (or more specifically from the fact that we take the error to follow a normal distribution).

3 Linear regression with fixed intercept [normal]

We can calculate $\hat{b}$ by minimizing the Sum of the Squared Errors (SSE).

$$SS_E = \sum_i (y_i - \hat{Y}_i)^2 = \sum_i (y_i - \hat{b}x_i)^2$$

$$\frac{\partial SS_E}{\partial \hat{b}} = \sum_i -2x_i(y_i - \hat{b}x_i) = -2\sum_i x_i y_i + 2\hat{b}\sum_i x_i^2)$$

Setting the derivative to 0 and solving for $\hat{b}$ we get :

$$\hat{b} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

Taking the second derivative we find

$$\frac{\partial^2 SS_E}{\partial \hat{b}^2} = 2\sum_i x_i^2 > 0$$

So it is a minimum (unless we have the pathological case where all $x_i = 0$ in which case we cannot define a slope).

Note that this result is NOT the same as starting form the estimator in the usual linear case. There, the "fixed point" is $(\bar{X}, \bar{Y})$, now it is $(0, 0)$.

4 Linear regression with a quadratic law passing by zero [advanced]

The estimator $\hat{b}$ can be found the same differential approach as used in the previous exercise.

$$SS_E = \sum_{i=1}^{10} (Y_i - \hat{b}x_i^2)^2 \qquad \frac{\partial SS_E}{\partial \hat{b}} = -2\sum_{i=1}^{10} (Y_i - \hat{b}x_i^2)x_i^2 = 0 \qquad \hat{b} = \frac{\sum_{i=1}^{10} x_i^2 Y_i}{\sum_{i=1}^{10} x_i^4} = 0.780$$

To check if $\hat{b}$ is non-biased, we have to prove its expected value is equal to b.

$$\mathbb{E}(\hat{b}) = \frac{\sum_{i=1}^{10} x_i^2 \mathbb{E}(Y_i)}{\sum_{i=1}^{10} x_i^4} = \frac{\sum_{i=1}^{10} x_i^2 b x_i^2}{\sum_{i=1}^{10} x_i^4} = b$$

So our estimator is unbiased.

As above, the parameters x_i are considered to be known and exact (i.e. they have no error). The errors we treat are in the y_i and expressed by ε_i. Since the average of a constant is the constant itself and since its variance is zero, $\mathbb{E}(Y_i) = bx_i^2$ and $\mathbb{V}(Y_i) = \mathbb{V}(\varepsilon_i) = \sigma^2$.

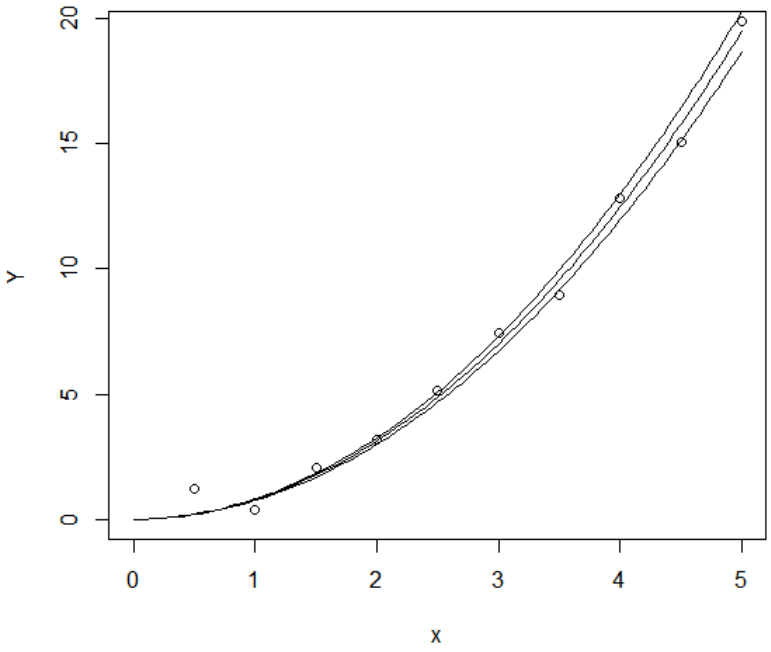
So $\hat{b}$ follows the normal law:

$$\mathcal{N}(\mathbb{E}(\hat{b}), \mathbb{V}(\hat{b})) = \mathcal{N}\left(b, \frac{\sum_{i=1}^{10} x_i^4 \mathbb{V}(Y_i)}{(\sum_{i=1}^{10} x_i^4)^2}\right) = \mathcal{N}\left(b, \frac{\sigma^2}{\sum_{i=1}^{10} x_i^4}\right)$$

And the 99% interval of confidence is (using 99.5 because we do a two-sided test, as above):

$$[b_i, b_s] = \left[\hat{b} \pm z_{99.5\%} \frac{\sigma}{\sqrt{\sum_{i=1}^{10} x_i^4}}\right] = [0.747, 0.812]$$

The figure below shows the linear regression as well as the functions with the upper and lower bound.



5 Efficient methods of learning a course and Model selection

To compare the different model, we will start from the "Full" model containing all contributions b_1, b_2, b_3 and see the effect of removing one of these slopes. To see the effect, we calculate the Fisher statistic using the following formula :

$$F_{obs} = \frac{(SS_{Efull} - SS_{Ereduced})/\Delta\nu_M}{SS_{Efull}/\nu_{Efull}}$$

where $\nu_{Efull} = n - p - 1$ is the error degree of freedom of the full model, with p the number of slopes (regressions) in the model. By only removing one slope at the time, we will always have $\Delta\nu_M = 1$.

This value is then compared to the critical Fisher Statistic $\mathcal{F}_{1,\nu_{Efull}}(P)$. If $F_{obs} > \mathcal{F}_{1,\nu_{Efull}}(P)$, then the reduced model (without the extra parameter) is statistically significantly worse than the full model. Conversely, if $F_{obs} < \mathcal{F}_{1,\nu_{full}}(P)$, the removed slope does not have a statistically significant effect on the full model. I.e. any lowering of the SSE that the full model provides could just be the result of random fluctuations (and the fact that with more free parameters, the fit is always better).

Starting with $a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3$ as the full model, with $\nu_{Efull} = 21 - 3 - 1 = 17$ and $\mathcal{F}_{1,17}(95\%) = 4.451$, we compare to the models with only 2 slopes and find (labelling them by the b parameter that was removed) :

$$F_{full-b1} = 28.2, \qquad F_{full-b2} = 12.4, \qquad F_{full-b3} = 0.95$$

So we can say with a confidence level of 95% that parameter b_3 does not contribute to a statistically significantly better full model, and removing it would be appropriate. In other words, we can't reject the hypothesis that $b_3 = 0$. However, we can reject the hypothesis that $b_2 = 0$ and even more so that $b_1 = 0$

To go further we now set $a + b_1 \cdot x_1 + b_2 \cdot x_2$ to be the $full'$ model, and look at the effect of removing 1 and 2. We now have $\nu_{full'} = 21 - 2 - 1 = 18$ and $\mathcal{F}_{1,18}(95\%) = 4.414$, we compare to the models with 1 slope and find :

$$F_{full'-b1} = 28.1, \qquad F_{full'-b2} = 12.4$$

So also here we find that we cannot remove any more parameters without significantly damaging the model.

Note that we do not compare the $full'$ model to the model with only b_3 as this would not be a nested model of the $full'$ model.