# Stats216v: Statistical Learning

Stanford University
Summer 2017
Gyu-Ho Lee (gyuhox@gmail.com (mailto:gyuhox@gmail.com))

## 3. Linear Regression

### *3.1.R1*

Why is linear regression important to understand? Select all that apply:

1. The linear model is often correct.
2. Linear regression is very extensible and can be used to capture nonlinear effects.
3. Simple methods can outperform more complex ones if the data are noisy.
4. Understanding simpler methods sheds light on more complex ones.

Gyu-Ho's Answer: 1, 3, 4.

2, 3, 4.

The linear model (and every other model) is hardly ever true, but it is an important piece in many more complex methods.

### *3.1.R2*

Which of the following are true statements? Select all that apply:

1. A 95% confidence interval is a random interval that contains the true parameter 95% of the time.
2. The true parameter is a random value that has 95% chance of falling in the 95% confidence interval.
3. I perform a linear regression and get a 95% confidence interval from 0.4 to 0.5. There is a 95% probability that the true parameter is between 0.4 and 0.5.
4. The true parameter (unknown to me) is 0.5. If I sample data and construct a 95% confidence interval, the interval will contain 0.5 95% of the time.

Gyu-Ho's Answer: 1, 3.

1, 4.

Confidence intervals are a "frequentist" concept: the interval, and not the true parameter, is considered random.

***3.2.R1***

We run a linear regression and the slope estimate is 0.5 with estimated standard error of 0.2. What is the largest value of $b$ for which we would NOT reject the null hypothesis that $\beta_1 = b$? (assume normal approximation to $t$ distribution, and that we are using the 5% significance level for a two-sided test; need two significant digits of accuracy)

Gyu-Ho's Answer: 0.892.

The 95% confidence interval $\hat{\beta}_1 \pm 1.96 * S.E.(\hat{\beta}_1) = 0.5 + 1.96 * 0.2 = 0.892$ contains all parameter values that would not be rejected at a 5% significance level. The critical value for a 95% confidence interval is 1.96.

***3.2.R2***

Which of the following indicates a fairly strong relationship between $X$ and $Y$?

1. $R^2 = 0.9$.
2. The p-value for the null hypothesis $\beta_1 = 0$ is 0.0001.
3. The $t$-statistic for the null hypothesis $\beta_1 = 0$ is 30.

Gyu-Ho's Answer: 1.

The $R^2$ is the correlation between the two variables and measures how closely they are associated. The $p$-value and $t$-statistic merely measure how strong is the evidence that there is a nonzero association. Even a weak effect can be extremely significant given enough data.

***3.3.R1***

Suppose we are interested in learning about a relationship between $X_1$ and $Y$, which we would ideally like to interpret as causal.

True or False: The estimate $\hat{\beta}_1$ in a linear regression that controls for many variables (that is, a regression with many predictors in addition to $X_1$) is usually a more reliable measure of a causal relationship than $\hat{\beta}_1$ from a univariate regression on $X_1$.

Gyu-Ho's Answer: False.

Adding lots of extra predictors to the model can just as easily muddy the interpretation of $\hat{\beta}_1$ as it can clarify it. One often reads in media reports of academic studies that "the investigators controlled for confounding variables," but be skeptical!

Causal inference is a difficult and slippery topic, which cannot be answered with observational data alone without additional assumptions.

***3.R.R1***

What is the difference between $lm(y \sim x * z)$ and $lm(y \sim I(x * z))$, when $x$ and $z$ are both numeric variables?

1. The first one includes an interaction term between x and z, whereas the second uses the product of x and z as a predictor in the model.
2. The second one includes an interaction term between x and z, whereas the first uses the product of x and z as a predictor in the model.
3. The first includes only an interaction term for x and z, while the second includes both interaction effects and main effects.
4. The second includes only an interaction term for x and z, while the first includes both interaction effects and main effects.

Gyu-Ho's Answer: 4.

An interaction term between a numeric x and z is just the product of $x$ and $z$. The difference is that in the first model, lm processes the "*" operator between variables and automatically includes main effects, whereas in the latter model, the expression inside the $I()$ function is not parsed as a part of the formula, but rather is simply evaluated.

***3.4.R2***

What is the predicted balance for an African American? (within .01 accuracy)

Gyu-Ho's Answer: Flexible is better, when we have much data.

A flexible model will allow us to take full advantage of our large sample size.

***3.Q***

Which of the following statements are true?



1. In the balance vs. income * student model plotted on slide 44, the estimate of beta3 is negative.
2. One advantage of using linear models is that the true regression function is often linear.
3. If the F statistic is significant, all of the predictors have statistically significant effects.
4. In a linear regression with several variables, a variable has a positive regression coefficient if and only if its correlation with the response is positive.

Gyu-Ho's Answer: 4.

1.

We can see that the estimate of beta3 is negative because the slope of the student line is smaller than the slope of the non-student line. That is, being a student diminishes the effect of income on balance.

The linear model is almost always wrong; however, it is often still useful.

The F statistic tests the null hypothesis that none of the predictors has any effect. Rejecting that null means concluding that *some* predictor has an effect, not that *all* of them do.

Positive correlation only means that the univariate regression has a positive correlation. See slide 20 for a counterexample.