Linear Regression Lines (Part 2)

Instructor: Hamed Hasheminia

Lecture 4

Agenda

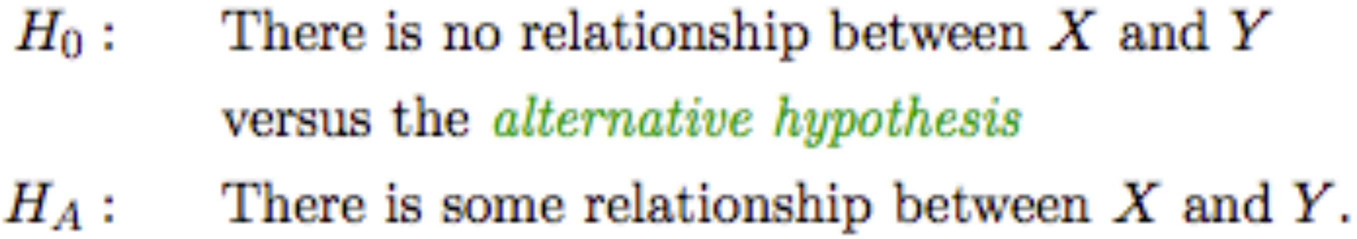
- Non-Linear Terms (Python)
- Hypothesis test – test of significance on regression coefficients
- P-values
- Different types of errors and R^2
- Interaction Effects

Hypothesis Test on Single Variable Regression lines

- Here is an example of a Single Variable Regression Model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

- What will happen to our coefficients if there is no relationship between X and Y? Can we still use X to predict Y?

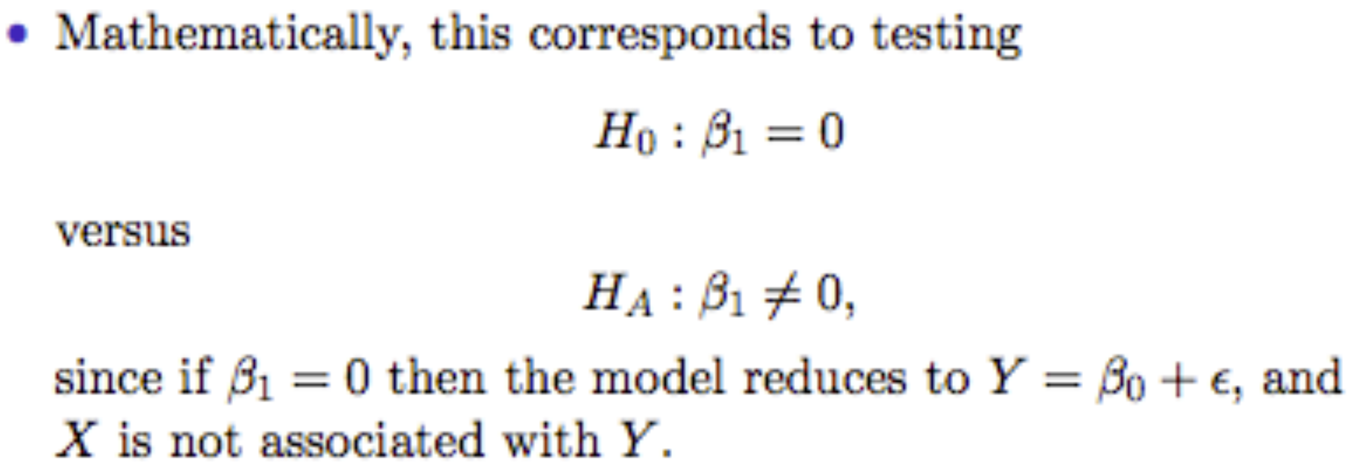


H_0 : There is no relationship between X and Y

versus the *alternative hypothesis*

H_A : There is some relationship between X and Y.

Hypothesis Test on Single Variable Regression lines



- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

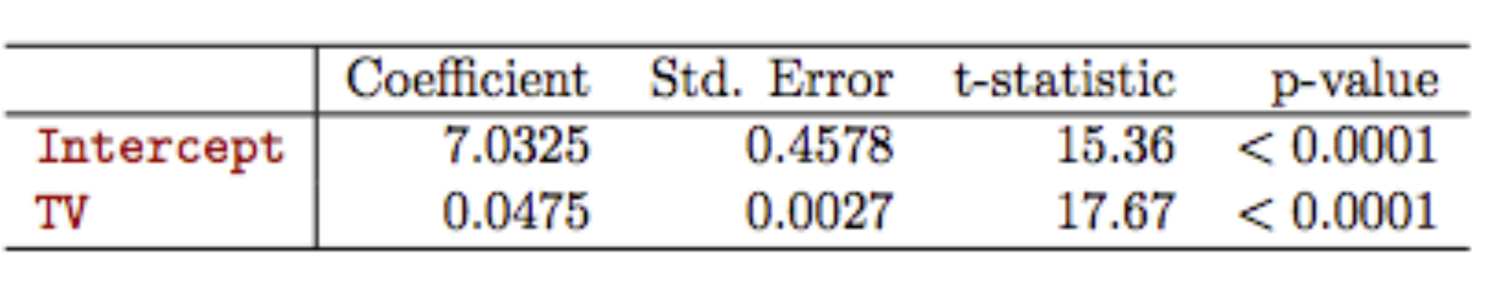
versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

Hypothesis Test on Single Variable Regression lines

- We use Statistical Software to compute the probability of estimating a coefficient as large as $\hat{\beta}_1$ given that true population coefficient is $\beta_1 = 0$.
- The above-mentioned probability is called the p-value.
- Ideally we want p-value be **as small as possible.**
- In many applications p-value < 5% is an acceptable value. In almost all application p-value < .1% is considered really good.



	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

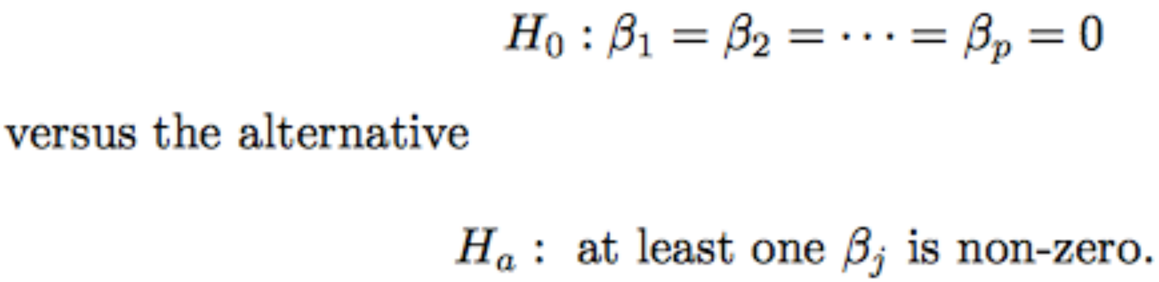
Hypothesis Test on Multi-Variable Regression lines

- For multi-Variable Regression lines we have two types of hypothesis test
  - Hypothesis test for significancy of our model
  - Hypothesis test for significancy of each of our variables.
  - Always start with Hypothesis test for the whole model and then check individual variables.
- Here is our multi-variable regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

- What values of coefficient will make our model useless?

Hypothesis test of the model



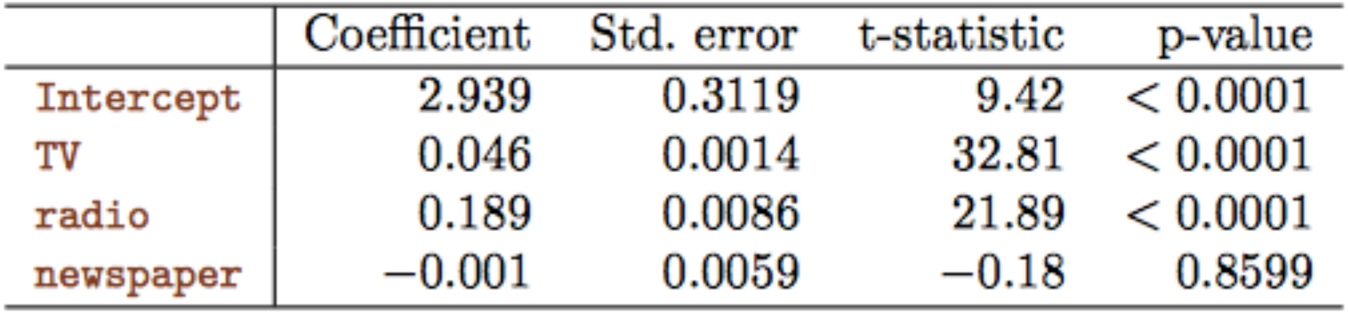
$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus the alternative

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

- Again we use statistical software to calculate the probability of estimating such large *beta_j*s given that null hypothesis is True. This is called the p-value of your model. Again, we want it be as small as possible. Usually p-values less than 5% is acceptable.

Hypothesis test of coefficients



	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

- The idea behind hypothesis test for coefficients is the same as single variable regression lines.
- You prefer to solely work with significant variables / If you observe insignificant variables you usually need to get rid of them and run your multivariable regression lines without those variables.
- If you have more than one insignificant variables, start dropping the most insignificant variable. If after removing that variable you still have insignificant variables, drop them one by one, until you are left with no insignificant variables. This method is called backward selection.

RSS, TSS, and R^2

- Our goal is to measure how much of the variability of the model is captured by our linear models.
- What is the worst model possible?
  - The worst model is something that does not use anything from your X variable to predict Y
  - My worst predictor is average value of Y. This value is independent from X. Any model shall perform better than this one.
  - Total Sum Square – TSS - is the error associated with our worst model. This is the error we would like to beat.
  - Residual Sum Square – is the error remained to be captured.
  - R^2 is the ratio of the captured error and total error.
  - R^2 = (TSS-RSS)/TSS = 1 – RSS/TSS

Interaction effects

- In our previous analysis of the Advertising data, we assumed that the effect on sales of increasing one medium is independent of the amount spent on the other media
- For example, the linear model

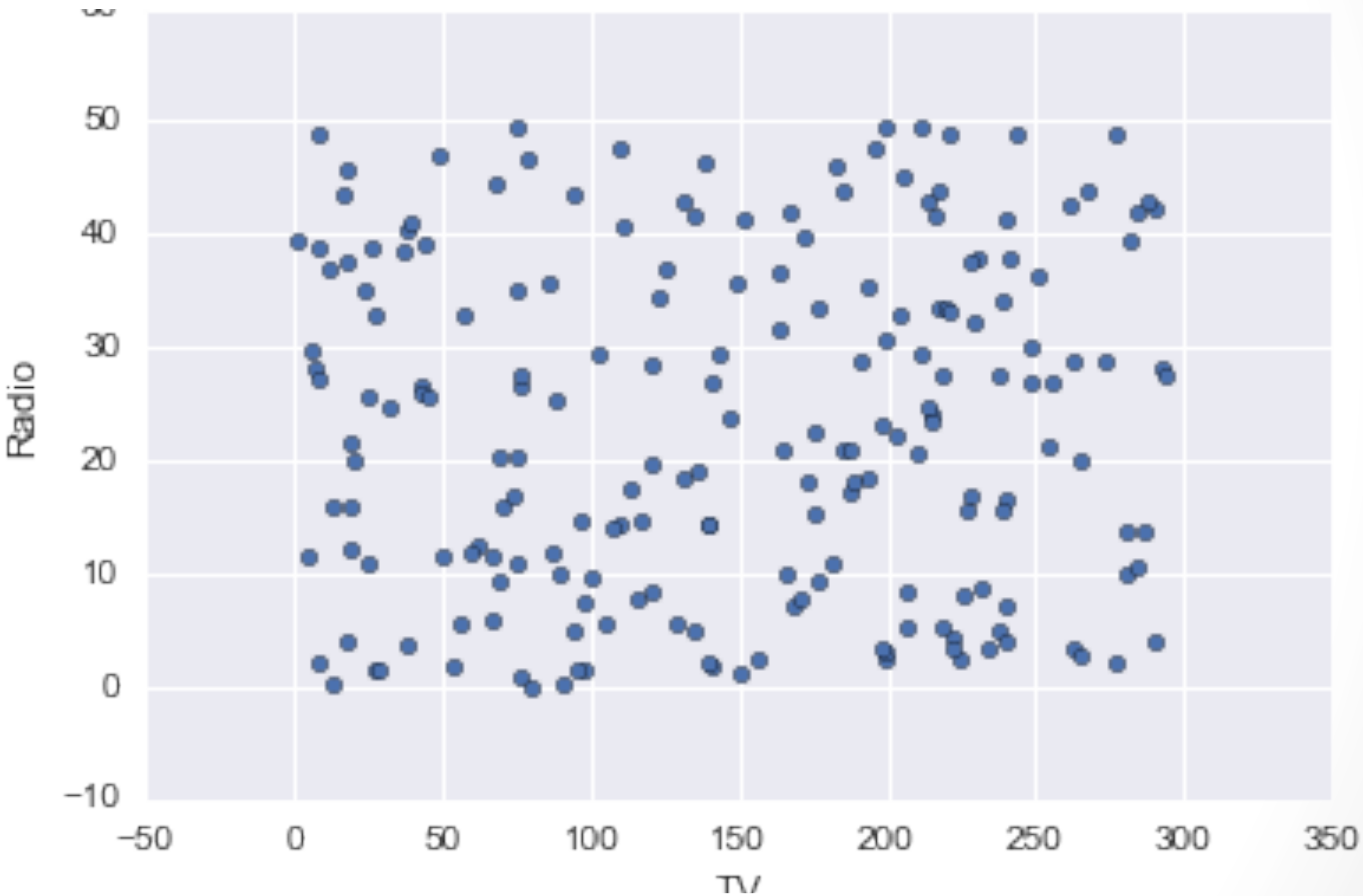
$$\widehat{\texttt{sales}} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$

states tat the average effect on sales of a one-unit increase in TV is always beta_1, **regardless of the amount spend on** radio.

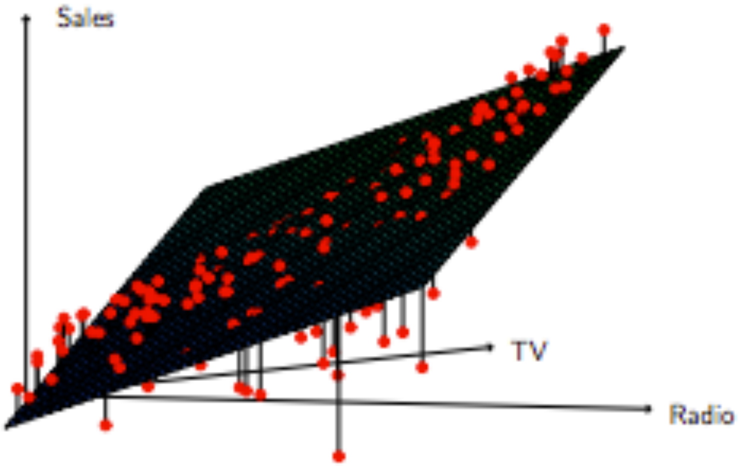
Interaction effects - continued

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.
- In this situation, given a fixed budget of $100,000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio.
- In marketing, this is known as a ***synergy*** effect, and in statistics it is referred to as an ***interaction*** effect.

TV vs Radio Scatterplot

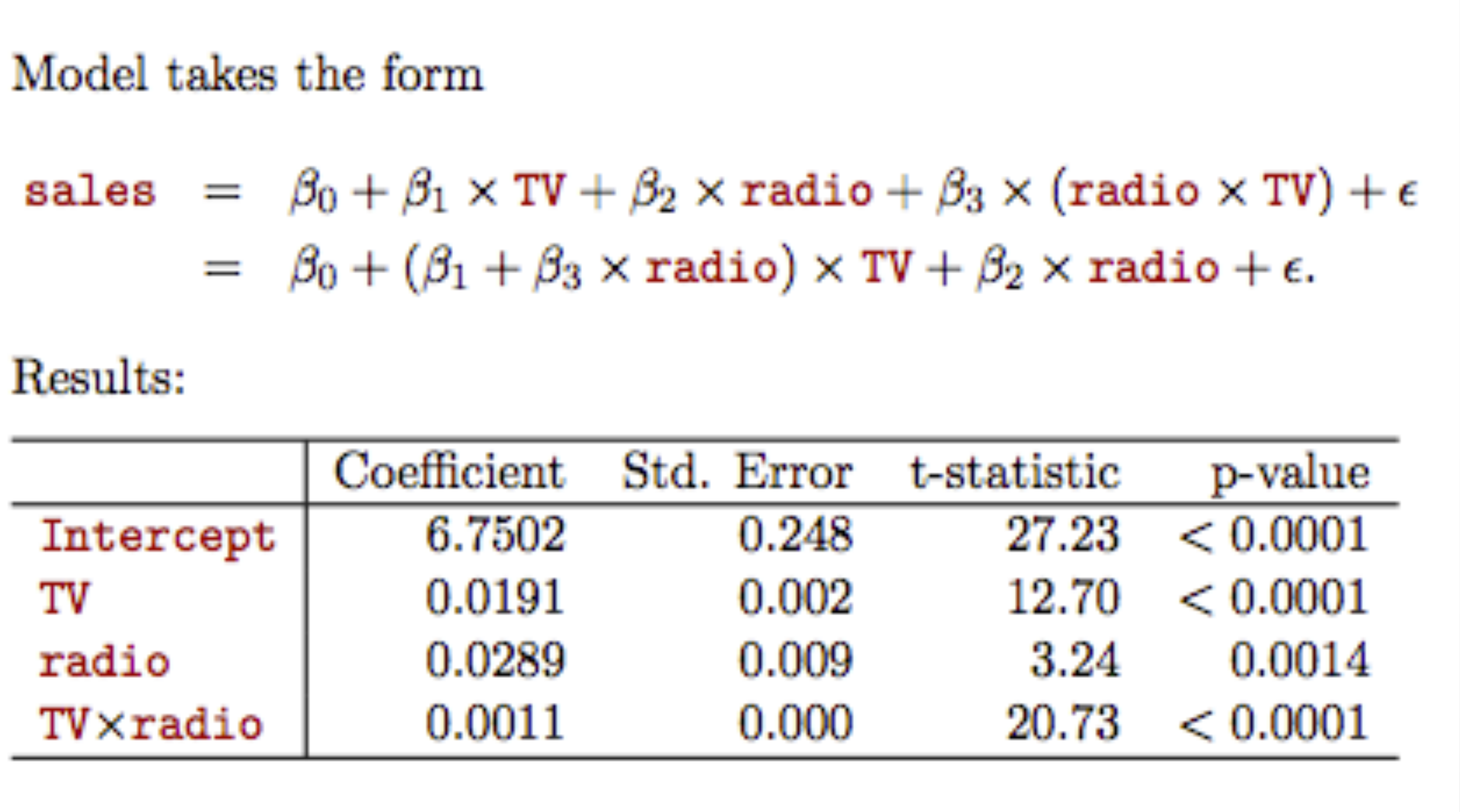


Interaction in the Advertising data?



- When levels of either TV or radio are low, then the true sales are lower than what predicted by the linear model.
- But when advertising is split between the two media, then the model tends to underestimate sales.

Modeling interactions – Advertising data



Model takes the form

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon.\end{aligned}$$

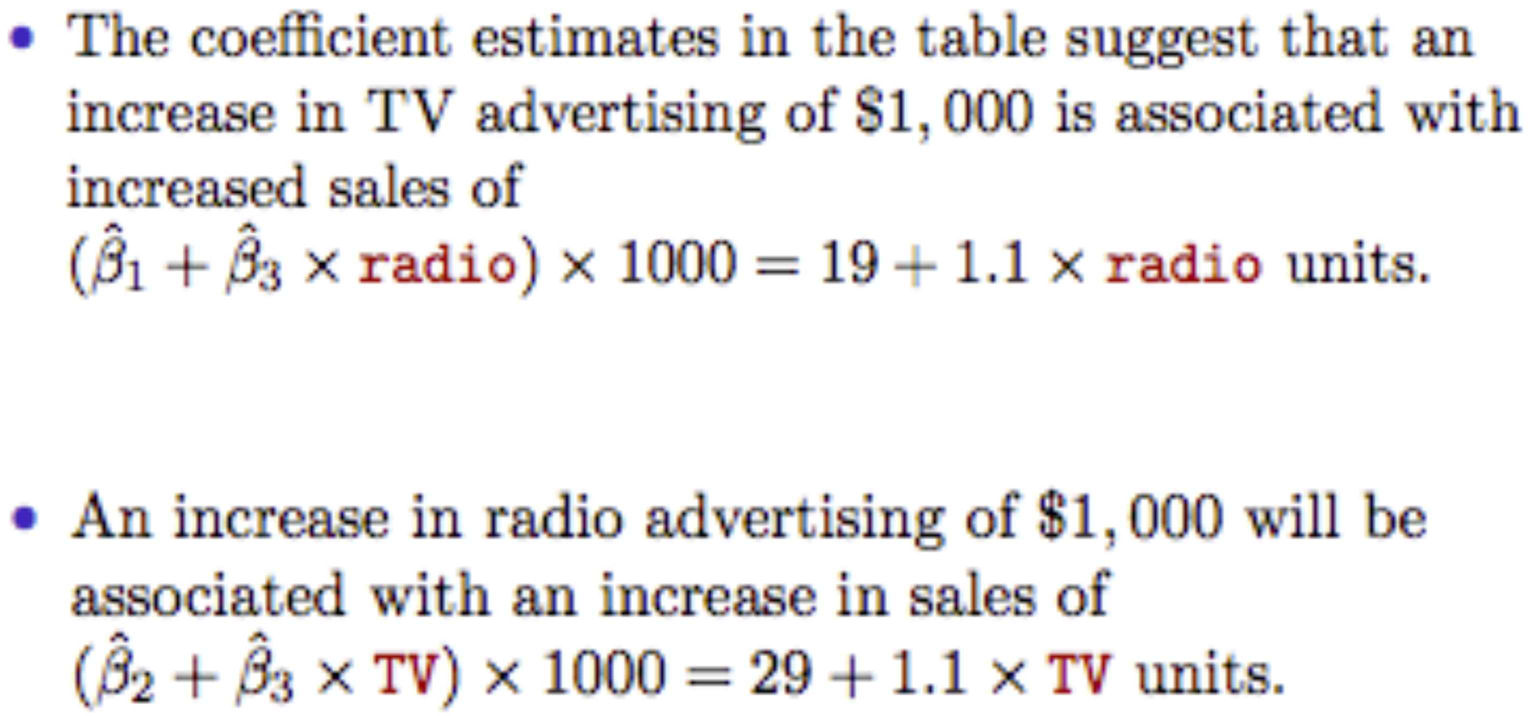
Results:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Interpretation

- The results in the table suggests that interactions are important.
- The p-value for the interaction term TV X radio is extremely low, indicating that there is strong evidence for beta_3 be different from zero.
- **The R^2** for the interaction model is 96.8%, compare to the 89.7% for the model without interaction. That means 69% of the variability in sales that remains after fitting the additive model has been explained by the interaction term.

Interpretation - continued



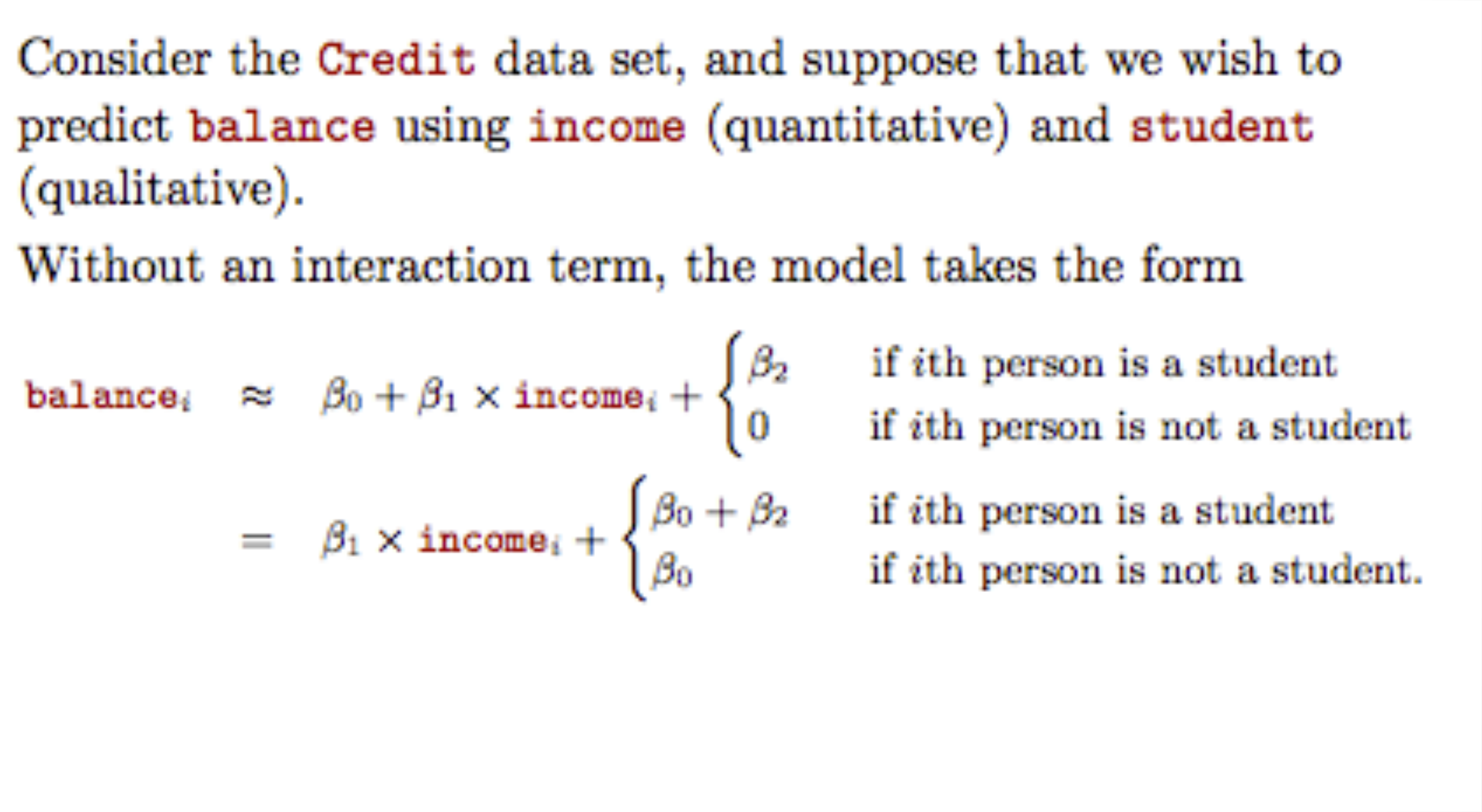
- The coefficient estimates in the table suggest that an increase in TV advertising of \$1,000 is associated with increased sales of $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.

- An increase in radio advertising of \$1,000 will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$ units.

Hierarchy

- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case TV and radio) do not.
- The ***hierarchy principle***
  - ***If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.***

Interaction between qualitative and quantitative variables

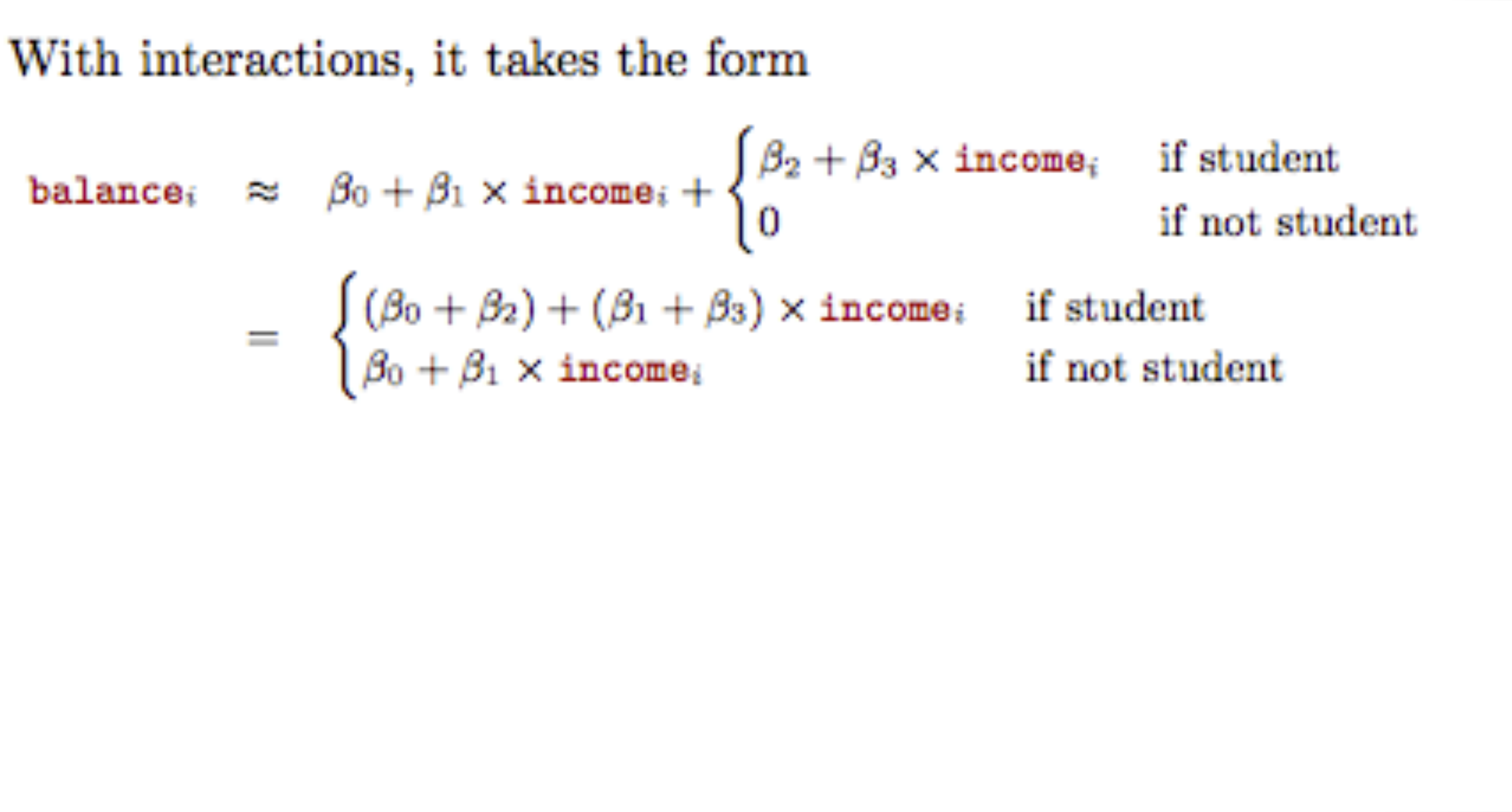


Consider the Credit data set, and suppose that we wish to predict balance using income (quantitative) and student (qualitative).

Without an interaction term, the model takes the form

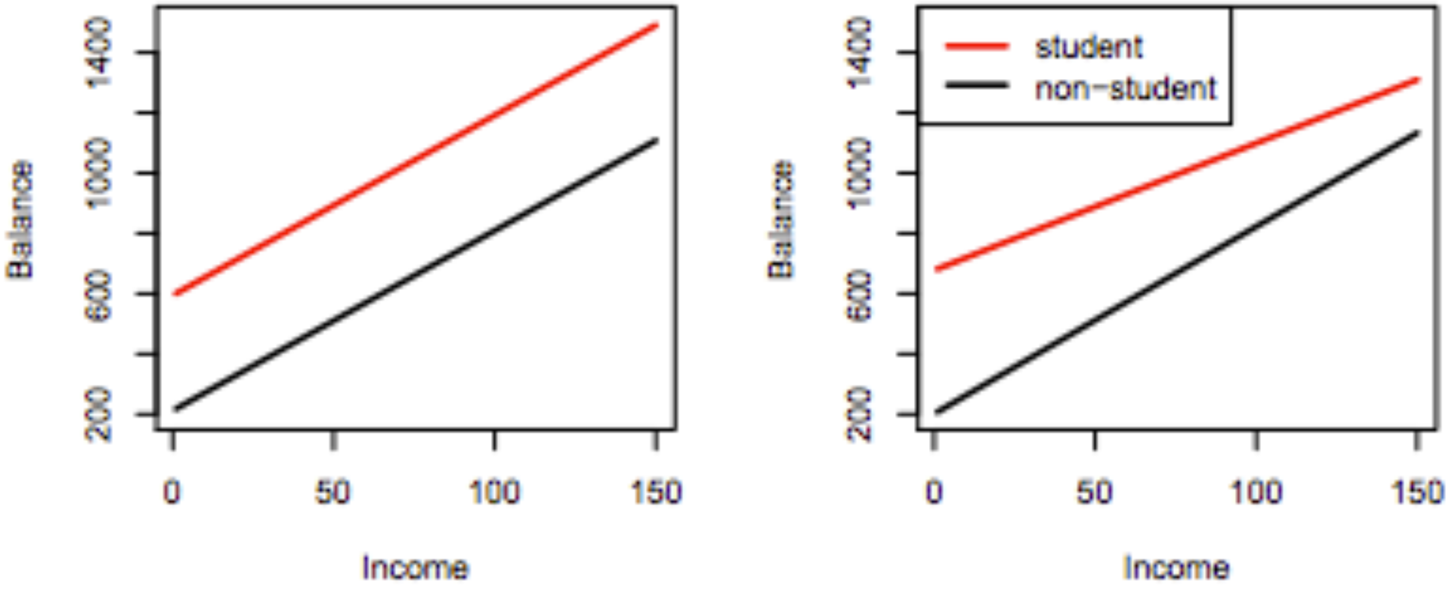
$$\begin{aligned}\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases}\beta_2 & \text{if } i\text{th person is a student}\\ 0 & \text{if } i\text{th person is not a student}\end{cases}\\ &= \beta_1 \times \texttt{income}_i + \begin{cases}\beta_0 + \beta_2 & \text{if } i\text{th person is a student}\\ \beta_0 & \text{if } i\text{th person is not a student.}\end{cases}\end{aligned}$$

Interaction between qualitative and quantitative variables



With interactions, it takes the form

$$\begin{aligned}\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases}\beta_2 + \beta_3 \times \texttt{income}_i & \text{if student}\\ 0 & \text{if not student}\end{cases}\\ &= \begin{cases}(\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student}\\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student}\end{cases}\end{aligned}$$



Credit data; Left: no interaction between income and student.
Right: with an interaction term between income and student.

Summary

- Hypothesis test on regression coefficients
- P-Value
- R^2
- Interaction effect (two quantitative variables)
- Interaction effect (A quantitative and categorical variable)