Multiple Linear Regression Review

Part II

Rebecca C. Steorts (slide adaption from Maria Tacket) and material from Chapter 1 of Roback and Legler text.

Computing set up

```
library(tidyverse)
library(tidymodels)
library(GGally)
library(xaringanExtra)
library(knitr)
library(patchwork)
library(viridis)
library(ggfortify)
library(dplyr)

ggplot2::theme_set(ggplot2::theme_bw(base_size = 16))

colors <- tibble::tibble(green = "#B5BA72")
```

Recap

Today's data is from the Kentucky Derby, an annual 1.25-mile horse race held at the Churchill Downs race track in Louisville, KY. The data is in the file derbyplus.csv and contains information for races 1896 - 2017.

Response variable

- `speed`: Average speed of the winner in feet per second (ft/s)

Additional variable

- `winner`: Winning horse

Predictor variables

- `year`: Year of the race
- `condition`: Condition of the track (good, fast, slow)
- `starters`: Number of horses who raced

Goal: Understand variability in average winner speed based on characteristics of the race.

Data

```
derby <- read_csv("data/derbyplus.csv")
# center the response by smallest response value
yearnew <- derby$year - min(derby$year)
#mutate(yearnew = derby$year - 1896)
derby <- cbind(yearnew, derby)
```

Data

```
derby |>
  head(5) |> kable()
```

yearnew	year	winner	condition	speed	starters
0	1896	Ben Brush	good	51.66	8
1	1897	Typhoon II	slow	49.81	6
2	1898	Plaudit	good	51.16	4
3	1899	Manuel	fast	50.00	5
4	1900	Lieut. Gibson	fast	52.28	7

Candidate models

Model 1: Main effects model (with centering)

```
model1Cent <- lm(speed ~ starters + yearnew +
                   condition, data = derby)
tidy(model1Cent) %>% kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	52.175	0.194	269.079	0.000
starters	-0.005	0.017	-0.299	0.766
yearnew	0.023	0.002	9.766	0.000
conditiongood	-0.443	0.231	-1.921	0.057
conditionslow	-1.543	0.161	-9.616	0.000

Candidate models

Model 2: Include quadratic effect for year

```
model2 <- lm(speed ~ starters + yearnew + I(yearnew^2)
             + condition, data = derby)
tidy(model2) %>% kable(digits = 3)
```

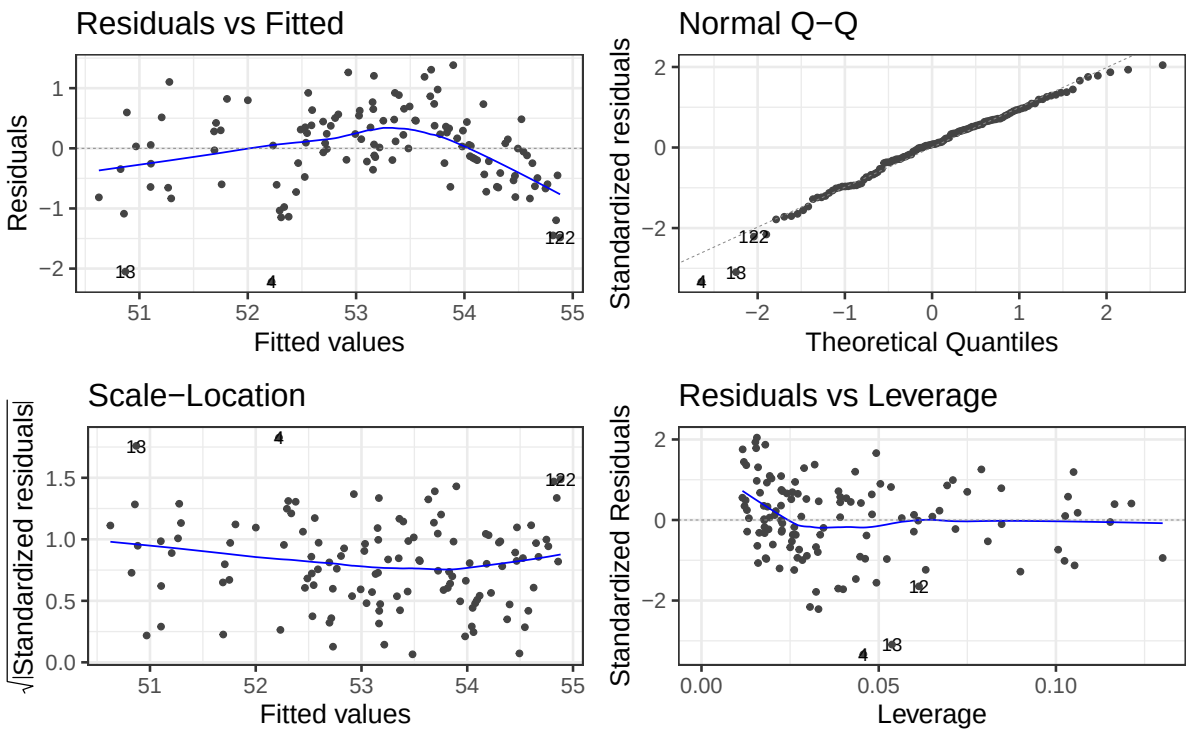
term	estimate	std.error	statistic	p.value
(Intercept)	51.413	0.183	281.564	0.000
starters	-0.025	0.014	-1.859	0.066
yearnew	0.070	0.006	11.424	0.000
I(yearnew^2)	0.000	0.000	-8.041	0.000
conditiongood	-0.477	0.186	-2.569	0.011
conditionslow	-1.393	0.131	-10.670	0.000

Model 1: Check model assumptions

1. The residuals versus fitted plot checks the linearity assumption. They should have no pattern around $Y = 0$. If not, this indicates a pattern in the data not accounted by the model.
2. The Normal Q-Q plot checks the normality assumption. Deviations from a straight line indicate the distribution of the residuals do not conform to a theoretical normal distribution.
3. The scale location plot checks the equal variance assumption. Positive/negative trends across the fitted values suggest the variability is not constant.
4. The residuals versus leverage plot is used to check for outliers (or highly influential points).

Model 1: Check model assumptions

```
autoplot(model1Cent)
```



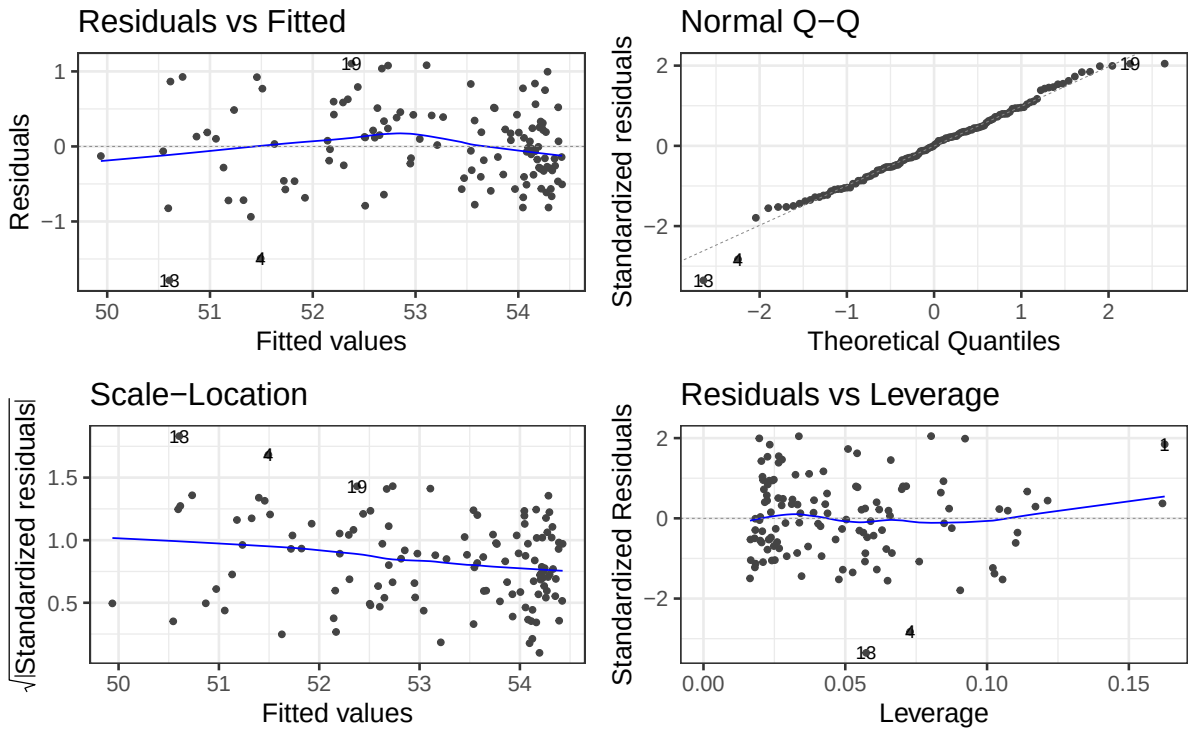
See BMLR, page 15 for a discussion and more details.

Model 1: Check model assumptions

- The residuals versus fitted values suggests a quadratic fit might be better than a linear one.
- The others appear reasonable.

Model 2: Check model assumptions

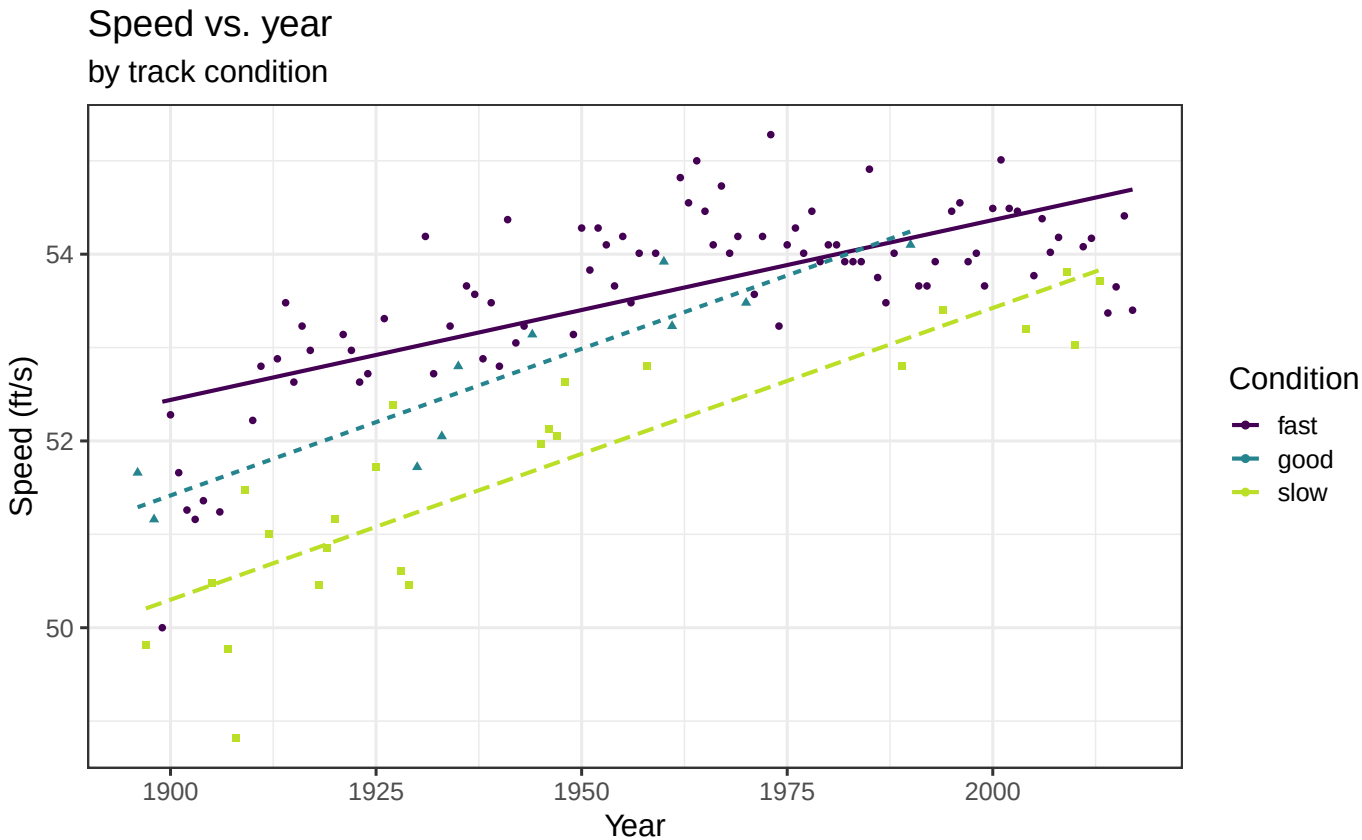
```
autoplot(model2)
```



Candidate models

What about an interaction term?

Recall from the EDA...



Model 3: Include interaction term

$$\begin{aligned}\widehat{speed} =& 52.387 - 0.003\ starters + 0.020\ yearnew - 1.070\ good \\ & - 2.183\ slow + 0.012\ yearnew \times good \\ & + 0.012\ yearnew \times slow\end{aligned}$$

term	estimate	std.error	statistic	p.value
(Intercept)	52.387	0.200	262.350	0.000
starters	-0.003	0.016	-0.189	0.850
yearnew	0.020	0.003	7.576	0.000
conditiongood	-1.070	0.423	-2.527	0.013
conditionslow	-2.183	0.270	-8.097	0.000
yearnew:conditiongood	0.012	0.008	1.598	0.113
yearnew:conditionslow	0.012	0.004	2.866	0.005

Interpreting interaction effects

term	estimate	std.error	statistic	p.value
(Intercept)	52.387	0.200	262.350	0.000
starters	-0.003	0.016	-0.189	0.850
yearnew	0.020	0.003	7.576	0.000
conditiongood	-1.070	0.423	-2.527	0.013
conditionslow	-2.183	0.270	-8.097	0.000
yearnew:conditiongood	0.012	0.008	1.598	0.113
yearnew:conditionslow	0.012	0.004	2.866	0.005

Measures of model performance

- R^2: Proportion of variability in the response explained by the model.
  - Will always increase as predictors are added, so it shouldn't be used to compare models
- $Adj.R^2$: Similar to R^2 with a penalty for extra terms

Measures of model performance

- *AIC*: Likelihood-based approach balancing model performance and complexity
- *BIC*: Similar to AIC with stronger penalty for extra terms

Measures of model performance

Nested F Test (extra sum of squares F test): Generalization of t-test for individual coefficients to perform significance tests on nested models

Which model would you choose?

Use the `glance` function to get model statistics.

model	r.squared	adj.r.squared	AIC	BIC
Model1	0.730	0.721	259.478	276.302
Model2	0.827	0.819	207.429	227.057
Model3	0.751	0.738	253.584	276.016

Which model would you choose?

Characteristics of a "good" final model

- Model can be used to answer primary research questions
- Predictor variables control for important covariates
- Potential interactions have been investigated
- Variables are centered, as needed, for more meaningful interpretations
- Unnecessary terms are removed
- Assumptions are met and influential points have been addressed
- Model tells a "persuasive story parsimoniously"

List from Section 1.6.7 of BMLR

Inference for regression

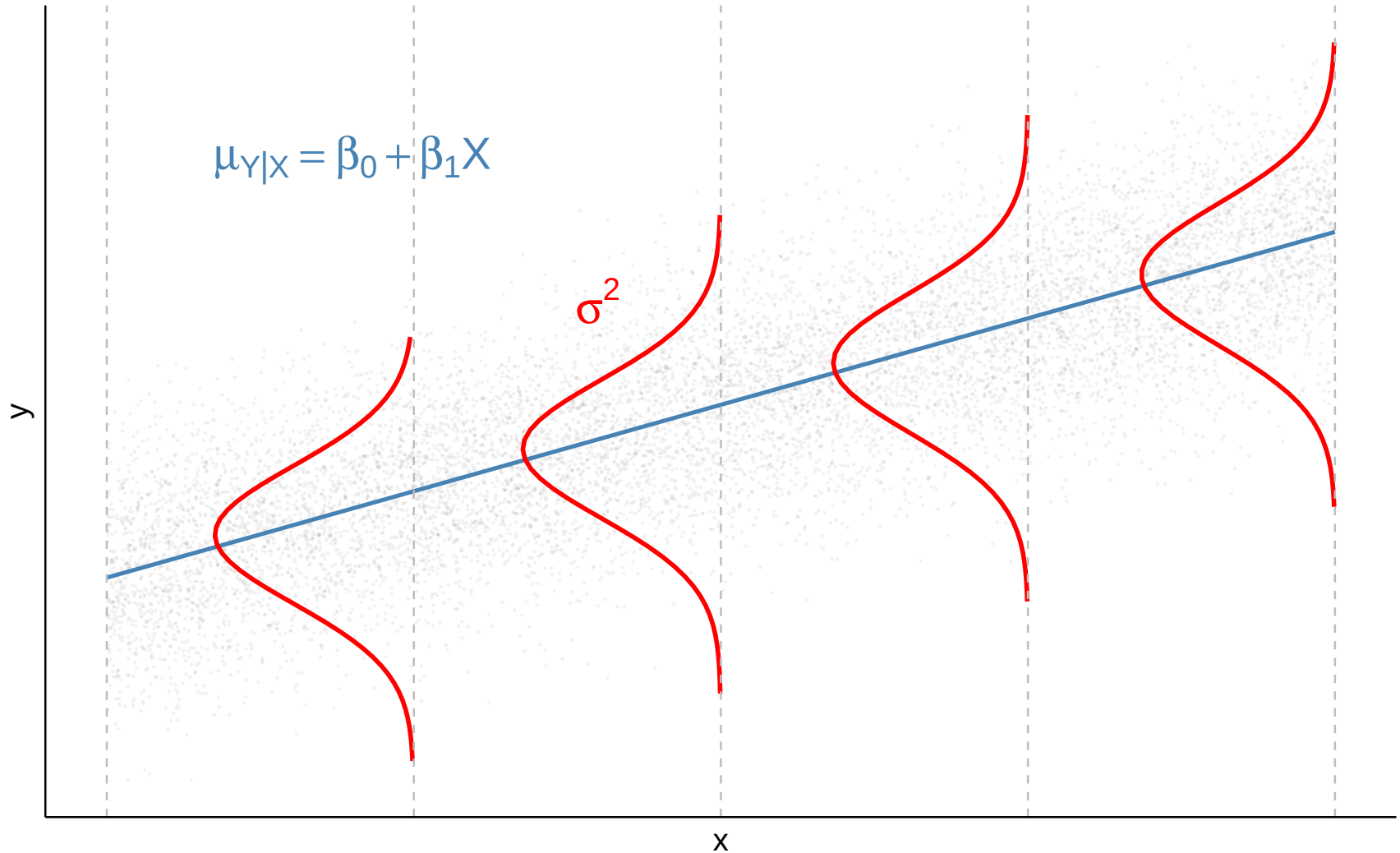
Use statistical inference to

- Evaluate if predictors are statistically significant (not necessarily practically significant!)
- Quantify uncertainty in coefficient estimates
- Quantify uncertainty in model predictions

If LINE assumptions are met, we can use inferential methods based on mathematical models. If at least linearity and independence are met, we can use simulation-based inference methods.

Inference for regression

When L.I.N.E. conditions are met



When L.I.N.E. conditions are met

- Use least squares regression to get the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$
- $\hat{\sigma}$ is the **regression standard error**

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-p-1}}$$

where p is the number of non-intercept terms in the model

(p $= 1$ in simple linear regression)

- Use $\hat{\sigma}$ to calculate $SE_{\hat{\beta}_j}$. Click here for more detail.

Inference for β_j

- Suppose we have the following model (for $i = 1, \ldots, n$):

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i, \quad \text{where} \\ &= \beta_0 + \sum_{j=1}^{p} x_{ji}\beta_j \\ \epsilon_i &\sim N(0, \sigma^2) \end{aligned}$$

Inference for β_j

- We use least squares regression to get estimates for the parameters $\beta_0, \beta_1, \ldots, \beta_p$ and σ^2. The regression equation is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- When the L.I.N.E. assumptions are met,

$$\hat{\beta}_j \sim N(\beta_j, SE_{\hat{\beta}_j})$$

- One objective of statistical inference is to understand β_j
- Use $\hat{\sigma}$ to estimate $SE_{\hat{\beta}_j}$, the **standard error of** $\hat{\beta}_j$

Inference for β_j

$$SE_{\hat{\beta}_j} = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_{x_j}^2}}$$

Conduct inference for β_j using a t distribution with $n - p - 1$ degrees of freedom (df).

- $\hat{\beta}_j$ follows a t distribution, because $\hat{\sigma}$ (not σ) is used to calculate the standard error of $\hat{\beta}_j$.
- The distribution has $n - p - 1$ df because we use up $p + 1$ df to calculate $\hat{\sigma}$, so there are $n - p - 1$ df left to understand variability.

Hypothesis testing for β_j

1. **State the hypotheses**. $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$, given the other variables in the model.
2. **Calculate the test statistic.**

$$t = \frac{\hat{\beta}_j - 0}{SE_{\hat{\beta}_j}}$$

3. **Calculate the p-value.** The p-value is calculated from a t distribution with $n - p - 1$ degrees of freedom.

$$\text{p-value} = 2P(T > |t|) \quad T \sim t_{n-p-1}$$

4. **State the conclusion in context of the data.**
   - Reject H_0 if p-value is sufficiently small.

Confidence interval for β_j

The $C\%$ confidence confidence interval for β_j is

$$\hat{\beta}_j \pm t^* \times SE_{\hat{\beta}_j}$$

$$\hat{\beta}_j \pm t^* \times \hat{\sigma} \sqrt{\frac{1}{(n-1)s_{x_j}^2}}$$

where the critical value $t^* \sim t(n-p-1)$

General interpretation: We are C percent confident that for every one unit increase in x_j, the response is expected to change by LB to UB units, holding all else constant.

Inference Activity (~8 minutes)

- Use the Model 3 output on the next slide to conduct a hypothesis test of the variable `yearnew` interpret the 95% confidence interval.
- You do not need to do the calculations by hand.

Model 3 output

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	52.387	0.200	262.350	0.000	51.991	52.782
starters	-0.003	0.016	-0.189	0.850	-0.035	0.029
yearnew	0.020	0.003	7.576	0.000	0.014	0.025
conditiongood	-1.070	0.423	-2.527	0.013	-1.908	-0.231
conditionslow	-2.183	0.270	-8.097	0.000	-2.717	-1.649
yearnew:conditiongood	0.012	0.008	1.598	0.113	-0.003	0.027
yearnew:conditionslow	0.012	0.004	2.866	0.005	0.004	0.020

Solution

1. **State the hypotheses**.
   $H_0 : \beta_{yearnew} = 0$ vs. $H_a : \beta_{yearnew} \neq 0$, given the other variables in the model.
2. **Calculate the test statistic.**

$$t = 7.576$$

3. **Calculate the p-value.**

The p-value is 0.

4. **State the conclusion in context of the data.**
   - Reject $H_{newyear}$ since the p-value is sufficiently small.

Solution

We are 95 percent confident that for every one unit increase in yearnew, the response is expected to change by 0.014 to 0.025 units, holding all else constant.